Propagation d'épidémies et survie en grande dimension : des modèles statistiques classiques aux méthodes d'apprentissage

Audrey Lavenu

Université de Rennes

Mémoire d'Habilitation à diriger des recherches

Propagation d'épidémies et survie en grande dimension : des modèles statistiques classiques aux méthodes d'apprentissage

Audrey LAVENU

IRMAR CNRS 6625

CIC 1414 INSERM

Soutenue le 14 avril 2023 devant le jury composé de :

- **Chantal Guihenneuc-Jouyaux**, Pr CNU 85, Université Paris Cité, France	Rapportrice
- **Agathe Guilloux**, Pr CNU 26, Université Paris Saclay, France en détachement DR INRIA, PariSanté Campus, France	Rapportrice
- **Julien Arino**, PhD, Professor at University of Manitoba, Winnipeg, Canada	Rapporteur
- **Laura Temime**, Pr CNU 26, CNAM, Paris, France	Examinatrice
- **Pierre Joly**, Pr CNU 26, Université Bordeaux, France	Examinateur
- **Valérie Monbet**, Pr CNU 26, Université de Rennes, France	Examinatrice Rennes
- **Magalie Fromont**, Pr CNU 26, Université Rennes 2, France	Garante de l'HDR

Avant propos

Je tiens à témoigner ma reconnaissance aux personnes sans qui ce travail n'aurait pu exister :

Magalie Fromont, rencontrée dès 2005 lors de mon année à l'agrocampus, la garante de ce travail, qui fut présente, motivante et exigeante, à chaque questionnement majeur dans ma carrière et dans ce travail de rédaction. Un clin d'œil à Burt qui l'a fait, merci pour tout.

Sandrine Katsahian, rencontrée il y a 23 ans lors de notre DEA de biomathématiques, tant d'anecdotes parisiennes et tant d'heures passées au téléphone, débouchant sur de nombreuses opportunités pour chacune séparément, mais aussi pour des projets communs qui commencent enfin à payer : début janvier 2023 notre premier (d'une longue série ;-)) papier accepté !

Valérie Garès, rencontrée lors de ce séminaire de stat à l'INSA il y a 5 ans, j'étais là au bon endroit au bon moment, sa volonté d'inclusion et son dynamisme furent le point de départ à mon rapprochement de l'IRMAR en 2018, merci de m'avoir poussée à plus de formalisme mathématique et pour la relecture de ce mémoire.

Muriel Hissler et **Eric Hitti** qui furent les déclencheurs de ma prise de décision de soutenir cette HDR lors des portes ouvertes de l'Université en février 2022.

Chantal Guihenneuc professeure statisticienne en section 85 (intervenantes toutes deux en 2009 dans le master de Jean-François Petiot à Vannes), **Agathe Guilloux** auteure d'un papier très inspirant pour le développement et la comparaison de méthodes d'analyse de survie en grande dimension, **Julien Arino** expert en modélisation des maladies infectieuses au Canada, merci beaucoup d'avoir accepté d'être rapportrices et rapporteur de ce travail et pour toutes les discussions qui en découlent.

Laura Temime, en DEA ensemble, en thèse ensemble, spécialiste entre autres, des modèles compartimentaux (on a pourtant jamais travaillé ensemble) et **Pierre Joly** de l'Université de Bordeaux spécialiste en survie, merci d'avoir accepté de faire partie de mon jury en tant qu'examinatrice et examinateur de mon HDR.

Valérie Monbet, responsable de l'équipe de stat de l'IRMAR, toujours facilitante (lors de mon CRCT, de ma demande de statut de chercheuse associée, de la venue de mon collègue anglais sur Rennes, etc.), merci d'avoir accepté d'être examinatrice de l'Université de Rennes dans mon jury.

Canelle Poirier et **Juliette Murris** mes deux premières expériences de co-encadrement de thèse, commencer avec deux chercheuses aux personnalités si agréables et efficacités impressionnantes, n'est pas offert à tout le monde, merci les filles.

Sylvain Duquesne et **Mihai Gradinaru**, ex et actuel directeurs de l'IRMAR, et **Jean-François Dupuy** ancien responsable de l'équipe de stat, qui m'ont accueillie et intégrée si chaleureusement depuis 2018. Bien sûr l'**IRMAR** en entier dont la tour de maths, Rennes 2, l'INSA et l'Agro, et tout particulièrement **Marie-Aude Verger** pour son professionnalisme administratif et sa gentillesse, **le groupe** *CHL/UFR maths/IRMAR et collègues présents dans les conseils/bureaux de l'Université de Rennes*, et la **commission parité**, qui m'apportent tant dans nos échanges de points de vue constructifs.

Eric Bellissant, qui m'a recrutée en 2006 au Centre d'Investigation Clinique (CIC) en m'impliquant dès le début dans le master MPCE puis dans la mention Santé Publique, et m'a mise en contact très

rapidement avec une liste de candidats à la Présidence de Rennes 1, créant l'opportunité de m'impliquer dans la politique universitaire depuis 15 ans. J'en profite donc également pour remercier mes **collègues vice-président(e)s**, **Guy Cathelineau** et **David Alis**, et les **services de l'Université** avec qui j'apprends beaucoup, et prends plaisir à travailler, échanger, construire, organiser, et apporter une petite pierre à l'édifice de projets ambitieux. Nous avons si souvent de nombreuses casquettes, et j'ai la chance de travailler avec certains et certaines sur plusieurs de nos champs de compétences, en particulier avec **Nicoletta Tchou** que je n'ai pas encore citée nominativement, qui est à l'IRMAR, dans le bureau du président, engagée sur la discrimination et l'égalité, et toujours disponible pour de bons conseils administratifs, politiques universitaires ou spécifiques à la recherche mathématique (dont le statut de chercheur associé :-)).

Pour continuer sur le **CIC**, le directeur actuel **Bruno Laviolle**, **Emmanuelle Comets** responsable de l'ex équipe pharmacocinétique PK et PK-PD, mes collègues de couloir, celles et ceux que je croise chaque année au couvent des jacobins, merci pour ces occasions d'encadrements de stagiaires, d'analyses statistiques diverses sur des sujets toujours si motivants, et les discussions autour d'un café. Le quotidien c'était aussi **Isabelle Merien** (Mam'Isabelle dixit certains de nos étudiants et mes enfants), quelle équipe on a fait sur le master 2 ;-), et **Delphine Bonnet** qui a pris la suite avec une grande rigueur, efficacité et investissement.

Catherine Houeix-Avril, qui a pris en charge quelques années la grosse partie du master 1 Santé Publique avec une implication énorme, un engagement direct auprès des étudiants et des enseignants, et le p'tit thé pour me faire tenir mon cours du mardi soir jusqu'à 19h30 ;-), **Yu Augagneur** qui a pris la suite et est toujours là pour nous faciliter la vie, et **Paula Molac** sur le master crimino, que de souvenirs également lors de sa construction en 2010 ! Il me faut également remercier l'ensemble des **collègues du Master 2 MPCE** où les réunions pédagogiques sont toujours agréables et enrichissantes depuis 16 ans, les **enseignants de l'UE Qualité et validation en 4ème année de pharma**, et le **labo d'informatique médicale** notamment **Marc Cuggia**, **Guillaume Bouzillé** et **Valérie Bertaud** avec qui le co-encadrement de Canelle Poirier fut un réel plaisir.

Guillaume Chauvet, dont le réseau anglais a permis ma mobilité à Southampton, et plus généralement l'**ENSAI** où mes cours de modélisation compartimentale en dernière année m'ont maintenu le pied à l'étrier sur cette thématique de recherche.

Samuel Jackson et **Dave Woods**, dont la collaboration initiée en juin 2019 à l'Université de Southampton, se maintient à distance en visio régulières, mais aussi en saisissant des opportunités de se voir à Southampton, à Rennes ou même en congés en Grèce ;-), merci de votre implication qui perdure malgré nos agendas d'enseignement chargés, notre travail va porter ses fruits :-). Merci également à l'ensemble des **collègues de stat et de maths** qui m'ont si bien accueillie avec des moments mémorables comme notre Random Walk sur l'île de Wight, et le port de la toge de cérémonie à la remise des diplômes.

Mes deux enfants qui ont souvent supporté l'ambiance de travail le soir et le week-end dès leur plus jeune âge, mais qui ont également pu partager cette magnifique expérience de vie que furent nos six mois à Southampton. **Leur père** qui m'a toujours soutenue depuis le début de mes études, notre quatuor qui se tire vers le haut et décuple les énergies, quel que soit le modèle structurel.

Ma sœur et **mes parents**, **mon oncle**, mais aussi **ma tante** qui aurait été si heureuse d'être présente pour la soutenance de cet ultime diplôme, **ma famille** entière, et **mes amis** avec qui tant de p'tits moments de bonheur permettent un bel équilibre.

J'espère n'avoir oublié personne parce que c'est **mon entourage complet professionnel et personnel** qui me permet d'avancer sereinement dans mes projets, et d'avoir pu prendre le recul nécessaire pour rédiger ce document qui m'a apporté bien plus qu'attendu en termes de prise de conscience de l'évolution de ma recherche, de liants insoupçonnés dus à mes appétences, de développements riches dans plusieurs domaines, et de motivation pour la suite.

Table des matières

Chapitre 1

Introduction

Mon profil de statisticienne a rencontré le domaine médical dès mon deuxième stage de fin d'études (celui de fin de maîtrise à l'INSERM), le DEA et la thèse de biomathématiques furent alors une évidence, et mes rattachements pluridisciplinaires aujourd'hui une conséquence logique : UFR Médecine, CNU de pharmacie (85, qualifiée en 26 et 67), affectation principale à l'IRMAR (unité mixte du CNRS en mathématiques), et un émargement secondaire au CIC1414 (Centre d'Investigation Clinique - INSERM).

Ce rattachement au Centre d'Investigation Clinique m'a permis d'apporter mes compétences statistiques dans de nombreux domaines pour répondre aux besoins des chercheurs cliniciens, vous retrouverez l'ensemble de mes papiers en annexe dans le chapitre 6 contenant une trentaine de travaux. Ce mémoire d'HDR se concentrera seulement sur les travaux concernant, de près ou de loin, la propagation d'épidémie et la survie en grande dimension (16 travaux [1]-[16]).

En plus des travaux avec plusieurs **chercheurs et chercheuses du CIC et de l'IRMAR**, j'ai pu développer des collaborations fructueuses avec :
- **l'équipe INRIA HeKA**, évolution de l'équipe « Science de l'information au service de la médecine personnalisée » de l'UMR 1138 - Centre de Recherche des Cordeliers (Paris Sorbonne - INSERM) des contrats précédents, au travers de mon deuxième co-encadrement de doctorante financée par une bourse CIFRE avec Pierre Fabre (avec un premier article accepté récemment sous réserve de révisions mineures [16]), mais aussi trois autres projets en cours (un article soumis sur les trajectoires de soins et la mortalité dans l'insuffisance cardiaque [17], le travail publié dans les JDS [14] et sa suite, le développement d'une adaptation d'un test de détection de rupture d'un processus de Poisson et son application à l'insuffisance cardiaque sur lequel travaille ma première doctorante en post-doctorat à l'Université de Rennes 2 avec Magalie Fromont),
- **le Laboratoire Traitement du Signal et de l'Image (LTSI - INSERM)**, au travers de mon premier co-encadrement de doctorante qui a conduit à 4 publications [11, 12, 13, 15].
- **le laboratoire de recherche de Harvard Medical School et Boston Children's Hospital**, au travers de la mobilité de ma première doctorante (articles [13, 15]),
- **l'équipe de statistique de l'Université de Southampton en Angleterre** (Southampton Statistical Sciences Research Institute) où j'étais en mobilité 6 mois en 2019, aboutissant à la publication d'un article dans une revue à fort impact sur l'asthme (article [18]), et un travail en cours sur l'identifiabilité des paramètres des modèles compartimentaux gérés par système d'équations différentielles).

Ce mémoire est le fruit de ces rencontres professionnelles riches et variées aux compétences si complémentaires et si inspirantes.

1.1 Les données de propagation d'épidémies

Une épidémie est définie par la propagation rapide d'une maladie contagieuse aux effets significatifs, touchant simultanément un grand nombre de personnes. Deux cas sont possibles : une augmentation d'une maladie endémique (c'est à dire où la présence de la maladie est permanente mais contenue à un taux constant dans une région ou une population particulière), ou l'apparition d'un grand nombre de malades là où il n'y avait rien avant. Le terme de pandémie n'a malheureusement aujourd'hui plus de secret pour personne. C'est une épidémie à l'échelle supérieure, quand le virus est capable de franchir les océans (sur une zone géographique très étendue en un temps assez court), quand il touche différentes sociétés, souvent la diffusion épidémique est alors d'un niveau plus grâve avec un plus fort taux de mortalité.
Pour illustrer ces définitions, prenons comme exemple la grippe chez l'homme. C'est une épidémie saisonnière dont les souches virales responsables sont de type B ou A (de sous-type A(H1N1) ou A(H3N2) faisant référence aux types de deux antigènes présents à la surface du virus : l'hémagglutinine de type 1 ou 3 et la neuraminidase de type 1 ou 2). Ces virus subissent des modifications mineures rapides appelées dérive antigénique.

La grippe espagnole, dont le nom n'est pas lié à l'origine ou l'ampleur épidémique mais seulement au fait que l'Espagne (n'étant pas en guerre) fut la seule à publier librement ses chiffres épidémiques, est également appelée « pandémie grippale de l'année 1918 ». C'est une pandémie de grippe A(H1N1) qui a débuté en mars 1918 et s'est propagée un peu plus d'un an et demi. Elle est responsable de 20 à 50 millions de morts selon l'Institut Pasteur, à cause d'une virulence importante d'une souche nouvelle non issue de la dérive antigénique des virus circulants, et d'une prévalence très grande de surinfections bactériennes fatales sans antibiotique (dont la découverte aura lieu 10 ans plus tard).

En avril 2009, nous étions dans un contexte de pandémie épizootique de grippe H5N1 qui affectait les oiseaux sauvages ou domestiques, avec un réservoir naturel entretenu par la migration. Certains volatiles avaient contaminé quelques hommes avec un taux de mortalité très fort, et faisait craindre une mutation du virus qui permettrait une transmission d'homme à homme par dérive antigénique et réarrangement génétique avec un virus grippal humain. Finalement, ce n'est pas la grippe aviaire H5N1 qui concrétisa le risque pandémique mais un nouveau virus de la grippe A de sous-type H1N1 qui contenait des gènes de plusieurs virus connus d'origines porcine, aviaire et humaine [19]. Ce virus sera appelé A(H1N1)pdm09, pour le différencier du sous-type A(H1N1) qui sévissait avant. Un an et demi après, le directeur de l'OMS déclarait : "Les pandémies sont de nature imprévisible et peuvent nous surprendre. Il n'y a pas deux pandémies semblables. Celle-ci s'est avérée beaucoup moins grave que nous avions pu le craindre il y a un peu plus d'un an. Cette fois-ci, nous avons eu beaucoup de chance. Le virus n'a pas muté pendant la pandémie vers une forme plus mortelle. Il n'est pas apparu de résistance généralisée à l'oseltamivir (l'antiviral approprié). Le vaccin s'est avéré bien adapté aux virus en circulation et son innocuité s'est révélée excellente...".

Les vaccins contre la grippe utilisés en France se composent de virus inactivés. Ils contiennent 4 souches de virus grippaux (2 souches de type A(H3N2+H1N1pdm09) et 2 souches de type B) qui sont choisies chaque année en fonction des souches qui ont circulé quelques mois avant dans l'hémisphère sud. Lors de ma thèse entre 2000 et 2004, j'ai pu participer à ces colloques et congrès spécifiques à la grippe, mais aussi au séquençage des virus grippaux de mon étude qui a enrichi la banque internationnale de séquences ARN.

En décembre 2019, une épidémie de pneumonies émerge dans la ville de Wuhan (province de Hubei, Chine). Le virus responsable est un coronavirus appelé SARS-CoV-2, différent du SARS-CoV responsable de l'épidémie de SRAS en 2003, qui avait à l'époque infecté environ 8 000 personnes dans une trentaine de pays, causant 774 décès. Il est également différent du virus MERS-CoV responsable d'une

épidémie évoluant depuis 2012 au Moyen-Orient (1219 cas diagnostiqués, 449 morts). Vous connaissez tous la suite...

La mesure principale de la propagation épidémique est l'incidence, définie par le nombre de nouveaux cas dans la population durant une période spécifiée de temps, ou plutôt le taux d'incidence (incidence rapportée à la population exposée en tenant compte de la durée de la période spécifiée).
Pour la grippe par exemple, le taux d'incidence chaque semaine est estimé à partir des déclarations de diagnostics de syndrômes grippaux des médecins de ville (donc basés sur les symptômes). Nous reviendrons dessus dans le chapitre 2. Pour la covid19, les indicateurs ont changé au fil du temps et sont aujourd'hui basés sur les résultats positifs des tests PCR, qui sont très dépendants des campagnes de dépistage et des recommandations officielles qui varient. Tout ceci doit être pris en compte dans les modélisations. J'évoquerai également ces données en fin de chapitre 3.

Sans déflorer le sujet, je caractériserais les méthodes utilisées pour étudier ce type de données en trois grandes classes : les séries temporelles, les méthodes d'apprentissage statistique et les modèles structuraux. Les deux premiers sont classiquement utilisés pour la prédiction bien qu'ils tendent aujoud'hui à répondre à des problématiques explicatives, nous verrons un peu plus loin dans l'introduction le débat engendré. Les modèles structuraux, que l'on restreindra ensuite aux modèles compartimentaux, devenus aujourd'hui populaires avec la notion de R0 qui doit être inférieure à 1 pour qu'une épidémie ne puisse démarrer, et la vulgarisation faite pour communiquer sur la COVID19. Les chapitres 2 et 3 présentront une grande partie des modèles de ces trois grandes classes, utilisés pour la propagation d'épidémies, nos applications sur la grippe, la gastro-entérite et la COVID19, et la recherche méthodologique mathématique et statistique que nous avons développée.

1.2 Les données de survenue d'évènement(s)

Les données de survenue d'évènements ou plutôt de durées jusqu'à ce qu'un événement se produise, comme la mort dans les organismes biologiques et la défaillance dans les systèmes mécaniques. Ce sujet est appelé théorie de la fiabilité ou analyse de la fiabilité en ingénierie, analyse de la durée ou modélisation de la durée en économie, et analyse de l'histoire des événements en sociologie. En médecine est souvent utilisé le raccourci *données de survie*, même si l'évènement n'est pas un décès, en anglais *Time-To-Event (TTE)* et *survival data*.

L'analyse de survie tente de répondre à certaines questions, telles que : quelle est la proportion d'une population qui survivra au-delà d'un certain temps ? Parmi ceux qui survivent, à quel rythme vont-ils mourir ou échouer ? Peut-on prendre en compte des causes multiples de décès ou d'échec ? Comment des circonstances ou des caractéristiques particulières augmentent-elles ou diminuent-elles la probabilité de survie ?

Un cas classique, qui complique le traitement des données, est la présence de risques compétitifs définis comme la situation où un autre événement se produit et s'oppose à la survenue de l'événement étudié (ou tout au moins altère fondamentalement la probabilité d'occurrence de l'événement d'intérêt).

Enfin, jusqu'ici nous avons évoqué les situations où un seul événement se produit pour chaque sujet, mais d'autres cas existent : les événements récurrents ou événements répétés, très courant dans les domaines de fiabilité des systèmes, en sciences sociales et dans la recherche médicale, je présenterai un travail dans ce cadre en fin de chapitre 4.

Les méthodes d'analyse de survie sont de plus en plus utilisées dans le domaine de l'oncologie. Pour obtenir des résultats fiables, le processus méthodologique et la qualité des rapports sont cruciaux. Une revue de la littérature intéressante argumente cette conclusion en étudiant les caractéristiques des analyses de survie dans les articles publiés dans les principales revues chinoises d'oncologie [20]. Le but était d'examiner la qualité méthodologique et celle du rapport de l'analyse de survie, d'identifier certaines déficiences communes, et de rédiger des recommandations. Un total de 242 articles d'analyse de survie ont été inclus pour être évalués parmi 1492 articles publiés dans 4 principales revues chinoises d'oncologie en 2013. Les articles ont été évalués selon 16 points établis pour une utilisation et un rapport appropriés de l'analyse de survie. Les taux d'application de Kaplan-Meier, de la table de survie, du test log-rank, du test de Breslow et du modèle des risques proportionnels de Cox (modèle de Cox) étaient respectivement de 91,74%, 3,72%, 78,51%, 0,41% et 46,28%, aucun article n'a utilisé la méthode paramétrique pour l'analyse de survie. Un modèle de Cox multivarié a été réalisé dans 112 articles (46,28%). Les violations et les omissions des directives méthodologiques comprenaient l'absence de mention des vérifications pertinentes de l'hypothèse de risque proportionnel ; l'absence de rapport sur les tests d'interaction et de colinéarité entre les variables indépendantes ; l'absence de rapport sur la méthode de calcul de la taille de l'échantillon. Trente-six articles (32,74 %) ont fait état des méthodes de sélection des variables indépendantes. Les défauts ci-dessus pourraient rendre potentiellement inexacts, trompeurs les résultats rapportés, ou difficiles à interpréter. Les auteurs de cette étude recommandent aux auteurs, aux lecteurs, aux rapporteurs et aux éditeurs de considérer l'analyse de survie avec plus d'attention et de coopérer plus étroitement avec les statisticiens et les épidémiologistes.

Comme dans tous les domaines, la complexité des données augmente et les données de survenue d'évènements n'y échappent pas. Là aussi la cancérologie n'est pas en reste au contraire, les protocoles de recherche clinique et les designs d'études sont très variables selon la phase de l'essai clinique (recherche de doses, pharmacodynamie, score de survie, mais aussi recherche de biomarqueurs que j'aborderai en début de chapitre 4).

Enfin, les données de génomique étant de plus en plus faciles à obtenir, elles ont un rôle croissant dans les modèles de survie qui ont la spécificité d'être appliqués à de petits échantillons de patients dans les essais cliniques. Les données de radiomiques peuvent aussi être intéressantes dans des problématiques de survie, et là encore sont de plus en plus courantes en cancérologie.

1.3 La grande dimension, une particularité des données massives en santé (big data in healthcare)

Pour schématiser la spécificité des données massives, nous pouvons citer les cinq « V » du big data [21] :
- Volume : la grande quantité d'information contenue dans ces bases de données.
- Vélocité : la vitesse de leur création, collecte, transmission et analyse.
- Variété : les différences de natures, formats et structures.
- Valeur : la capacité de ces données à générer du profit.
- Véracité : leur validité, i.e. qualité et précision ainsi que leur fiabilité.

Dans le domaine de la santé, l'importance de ces cinq dimensions diffèrent forcément par rapport à d'autres domaines tel que le marketing, mais avoir en tête ces 5 « V » permet d'être vigilant sur la qualité des données et l'impact éventuel en fonction du mode de recueil.

Par exemple, les données du Système National des Données de Santé (SNDS) qui est géré par la Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés (CNAMTS), sont basées sur les rembourse-

ments. Le Volume est énorme, la Véracité plutôt bonne, les informations sont fournies en fonction de la problématique de remboursement et non des questions cliniques, et nécessite beaucoup de travail sur les données pour s'en servir en épidémiologie.

Dans le chapitre 2, je décrirai l'entrepôt hospitalier eHOP, qui lui, comprend une grande "Variété" avec tous les documents scannés de type compte-rendus, résultats biologiques, etc. avec d'autres difficultés, enfin nous utiliserons également à plusieurs reprises des données comptabilisant des requêtes d'internautes sur Google ou des posts sur Twitter, qui là aussi nécessitent une vigilance accrue pour s'assurer de collecter le bon signal.

Enfin, les *données en grande dimension* sont un cas particulier des données massives, elles se caractérisent par un très grand nombre de covariables et moins de sujets. En recherche clinique avec des problématiques de survie, ce cas là se rencontre souvent, comme évoqué précédemment. Les conséquences sur les méthodes d'analyse sont très grandes et la recherche dans ce domaine en plein essort [22]. Nos études de simulation du chapitre 4 se feront dans ce contexte de grande dimension.

1.4 Le débat "modèles explicatifs versus prédictifs" qui décloisonne les méthodes d'analyse

Ces quelques lignes sont issues de la présentation de Gilbert Saporta intitulée "Expliquer ou prédire? Les nouveaux défis" [23]. La modélisation statistique est un outil puissant pour développer et tester des théories par le biais de l'explication causale, de la prédiction et de la description. Dans de nombreuses disciplines, est utilisée presque exclusivement la modélisation statistique pour l'explication causale avec l'idée que les modèles à fort pouvoir explicatif ont par nature un fort pouvoir prédictif. La confusion entre l'explication et la prédiction est courante, mais la distinction doit être comprise pour faire progresser la connaissance scientifique.
La modélisation explicative se base sur une théorie sous-jacente et un ensemble restreint de modèles, qui découlent de cette théorie, et seront testés, la qualité d'ajustement se fait sur la prédiction du passé, et l'erreur est un bruit blanc ; alors que **la modélisation prédictive** se base, elle, sur des modèles issus des données, utilise des modèles algorithmiques pour prédire de nouvelles données (ou l'avenir), et l'erreur est à minimiser.

Opposer ces deux cultures (très bien décrites dans l'article [24]) conduit à un vrai paradoxe de comprendre sans prédire avec un « bon » modèle qui s'ajuste bien mais peut fournir des prévisions médiocres au niveau individuel, et de prédire sans comprendre avec des modèles ininterprétables mais qui peuvent donner de bonnes prévisions. Pour rajouter une couche de confusion, dans certaines conditions, il a été montré que le « vrai » modèle ne prédit pas toujours mieux [25].

Par ailleurs, l'argument "comprendre pour mieux prédire" est associé à une confusion entre corrélation et causalité. Il est toujours très difficile d'inférer la causalité à partir de données d'observations. Un challenge important est de réaliser de l'inférence causale avec les données massives [26, 27].

L'article "Machine Learning in Medicine : To Explain, or Not to Explain, That Is the Question" [28] montre que la question reste d'actualité et compare les deux approches dans un contexte de machine learning en médecine.

Dans les chapitres 2 et 4, où nous serons dans le cadre de la grande dimension, notre objectif sera

la modélisation prédictive, avec une étude rigoureuse des performances prédictives. Nous analyserons néanmoins l'importance des variables, en gardant en tête ce débat, et comparerons les modèles statistiques utilisant des régressions pénalisés avec des méthodes d'apprentissage supervisé.

Pour finir cette introduction

Je donne les grandes lignes du plan de ce mémoire qui présentera les modèles mathématiques et statistiques utilisés, ou développés pour répondre à certaines problématiques cliniques liées de près ou d'un peu plus loin à la propagation d'épidémie ou la survie en grande dimension (articles [1]-[16]). Plusieurs de mes travaux sont basés sur des études de simulation pour évaluer des performances de modèles, d'autres sur la construction mathématique de modèles, et d'autres encore sur des propriétés statistiques d'identifiabilité de paramètres dans des systèmes d'équations différentielles. Les chapitres 2 et 3 concernent la propagation d'épidémie, le chapitre 2 en termes de modélisation de séries temporelles dans un cadre de grande dimension, et le chapitre 3 en termes de modélisation compartimentale SIR (Susceptibles-Infectious-Removed). Le chapitre 4 concerne la modélisation des données de survenue d'évènement(s) dans un cadre de grande dimension. Ces trois chapitres se terminent par des parties "Perspectives" pour expliquer mes projets à court et plus long termes, suivis d'un résumé de mes contributions à ces thématiques.

Chapitre 2

Propagation d'épidémie : modélisation de séries temporelles dans un cadre de grande dimension

Ce chapitre repose sur 4 articles[1] de la thèse de Canelle Poirier que j'ai co-encadrée, intitulée "Modèles statistiques d'aide à la décision en santé publique basés sur la réutilisation des données massives en santé : application à la surveillance syndromique.".
En redonnant les bases des modèles classiques en séries temporelles, je présenterai les modèles statistiques construits pour l'utilisation des données massives dans l'analyse de séries temporelles publiés dans trois de ces articles. La soutenance de la thèse s'est déroulée le 13 juin 2019 devant un jury pluridisciplinaire d'experts en statistique, en sciences des données à Santé Publique France, et en informatique médicale. En particulier, Pr Eric Matzner-Lober était un des deux rapporteurs, Pr Magalie Fromont était membre du comité de suivi de thèse. Ce travail s'inscrivait dans le cadre du projet **ANR INSHARE** (2016-2019 INtegrating and SHAring health data for REsearch) porté par un consortium d'équipes du LTSI, de LATIM/Télécom bretagne, de l'ENSAI, et de l'EHESP, et a intégré une mobilité de 6 mois de la doctorante à Harvard University bénéficiant du financement très sélectif **Fulbright** U.S. Student Program.

2.1 Modélisation de séries temporelles par méthodes dites traditionnelles

Considérons qu'une série temporelle observée $\{y_t, t = 1, \ldots, T\}$ est la réalisation de variables aléatoires $\{Y_t, t = 1, \ldots, T\}$ qui forment elles-mêmes une portion d'un processus aléatoire $\{Y_t, t = \ldots, 0, 1, 2, \ldots\}$, c'est-à-dire d'une série infinie de variables aléatoires. Avant de parler de modèle, et pour visualiser une caractéristique naturelle des séries temporelles, nous parlerons également d'auto-corrélation de la série qui fait référence au fait que la mesure d'un phénomène à un instant t peut être corrélée aux mesures précédentes (au temps $t-1, t-2, t-3, \ldots$) ou aux mesures suivantes (à $t+1, t+2, t+3, \ldots$). Une série auto-corrélée est ainsi corrélée à elle-même, avec un décalage (lag) donné. Et la fonction d'auto-corrélation

1. Pour faciliter la lecture de ce chapitre le 2ème article sera surnommé "Article Grippe uni-source", le 3ème "Article Grippe multi-sources" et le 4ème "Article Gastro bi-sources".

s'écrit : $\rho_h = \frac{\gamma(h)}{\gamma(0)}$ où $\gamma(h) = cov(Y_t, Y_{t+h})$.

2.1.1 Modèles autorégressifs

Modèle autorégressif d'ordre p. Le modèle AutoRégressif d'ordre p (AR(p)) s'écrit :

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t, t \in \mathbb{Z}, \epsilon_t \sim BB(0, \sigma_\epsilon^2).$$

où β_0 est une constante et $\phi_1, \ldots, \phi_p$ p coefficients, $\{\epsilon_t\}$ est un bruit blanc (BB) défini comme une suite de variables aléatoires non corrélées de moyenne nulle et de variance constante σ_ϵ^2 et $\{Y_t\}$ stationnaire (ou tout au moins "faiblement stationnaire").
Pour vérifier cette dernière condition, posons $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ le polynôme caractéristique de l'équation de récurrence qui décrit un AR(p) où B est l'opérateur retard.
Le modèle s'écrit alors :

$$\Phi(B) Y_t = \beta_0 + \epsilon_t$$

Si les racines de l'équation $\Phi(B) = 0$ sont en module > 1, la condition de stationnarité est vérifiée. Nous utilisons le test de Dickey-Fuller augmenté pour tester l'hypothèse nulle de non stationnarité, basé sur la présence de racines unitaires.
Pour l'hypothèse de bruit blanc, nous utilisons le test de Box-Pierce pour tester l'hypothèse nulle d'absence d'auto-corrélation des résidus.

Modèle autorégressif moyenne mobile. En rajoutant une partie "moyenne mobile" au modèle précédent, nous obtenons un modèle AutoRegressive Moving Average (ARMA(p,q)) qui s'écrit :

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q},$$

où $\epsilon_t \sim BB(0, \sigma_\epsilon^2)$ et $\theta_1, \ldots, \theta_q$ q coefficients.
Le modèle ARMA particulier sans constante ($\beta_0 = 0$) peut alors s'écrire avec B l'opérateur retard

$$\Phi(B) Y_t = \Theta(B) \epsilon_t$$

où $\Theta(B)$ s'écrit de la même manière que $\Phi(B)$ avec le paramètre q : $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_p B^q$, et donc

$$Y_t = \frac{\Theta(B)}{\Phi(B)} \epsilon_t$$

Modèle autorégressif moyenne mobile intégré. Le I de ARIMA est pour « integrated » et indique qu'il faut différencier la série originale afin d'éliminer un caractère non-stationnaire éventuel en travaillant sur $(Y_t - Y_{t-1})$.
En plus des paramètres p (nombre de décalages à considérer pour le modèle auto-régressif) et q (l'ordre du modèle MA, partie moyenne mobile), est rajouté un troisième paramètre (le nombre de fois qu'il faut différencier la série afin de la rendre stationnaire), qui sera donc fixé à 0 dans le cas d'un processus déjà stationnaire correspondant à un modèle ARMA(p,q).

Modèle autorégressif moyenne mobile avec composante explicative X. Le modèle ARMAX comporte donc une composante explicative X et une erreur ARMA (respectivement ARIMA pour ARIMAX), il s'écrit :

$$Y_t = \beta_0 + X_t \beta + u_t, \quad u_t = \frac{\Theta(B)}{\Phi(B)} \epsilon_t, \quad \epsilon_t \sim BB(0, \sigma_\epsilon^2).$$

2.1.2 Estimation

Identification. Pour ajuster un modèle AR ou ARMA à des données réelles, il faut d'abord choisir p et q du modèle ARMA (AR peut en effet être considéré comme un cas particulier de modèle ARMA avec q=0). Pour un modèle AR(p) p sera choisi tel que $h \geq p+1$ où la fonction d'autocorrélation $\rho(h)$ décroît vers 0 et la fonction d'auto-corrélations partielle s'annule r(h)=0. Pour un modèle ARMA(p,q) q sera choisi tel que $h \geq q+1$ où la fonction d'autocorrélation $\rho(h)$ décroît vers 0, et p tel que $h \geq \max(q+1, p+1)$ où la fonction d'auto-corrélations partielle r(h) décroît vers 0. Il n'est pas toujours facile de trouver p et q de cette manière, et un ensemble de modèles ARMA(p,q) sont parfois départagés par les critères AIC ou/et BIC.

Estimation. Plusieurs méthodes d'estimation des paramètres ϕ et θ sont possibles, celle des moindres carrés conditionnels est courante et par défaut sous plusieurs logiciels. Supposons $y_1, \ldots, y_p$ fixés et connus et que $\epsilon_p = \epsilon_{p+1} = \ldots = \epsilon_{p+q} = 0$, alors pour un processus ARMA(p,q), par récurrence

$$\epsilon_t = y_t - \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}.$$

La somme des carrés conditionnels aux valeurs initiales s'écrit :

$$CSS = \sum_{t=1}^{T} \epsilon_t^2 = \sum_{t=1}^{T} [y_t - \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}]^2,$$

où les ϵ_{t-j} peuvent être écrits en fonction des $y_{t-j}, \ldots, y_{t-j-p}$ et des $\epsilon_{t-j-1}, \ldots, \epsilon_{t-q}$. Le critère CSS sera minimisé par algorithme itératif pour estimer ϕ et θ.

2.1.3 Application aux données grippales et de gastro-entérite françaises

Aujourd'hui, à travers le monde, la surveillance sanitaire est au cœur des enjeux de santé publique. La pandémie COVID-19 en est l'illustration qui surpasse toute démonstration connue depuis longtemps des risques de propagation des maladies [29]. L'OMS se charge de coordonner la prévention et la coordination entre les pays pour une veille sanitaire efficace au travers de réseaux d'information. En France, c'est Santé publique France, l'agence du ministère de la Santé qui se définit ainsi :
"Par la veille et la surveillance épidémiologiques, l'agence anticipe et alerte. Par sa maîtrise des dispositifs de prévention et de préparation à l'urgence sanitaire, elle accompagne les acteurs engagés de la santé publique. Ancrée dans les territoires, elle mesure l'état de santé et déploie ses dispositifs au plus près des publics, dans un souci constant de fonder une connaissance juste et de proposer des réponses adaptées." [30].
Santé publique France s'appuie sur le réseau national de santé publique qui regroupe d'un côté, les réseaux de veille et de surveillance, et de l'autre, les réseaux de prévention et de promotion de la santé, comme par exemple les Agences Régionales de la Santé (ARS) et le réseau Sentinelles [31]. La surveillance syndromique a été définie par le Center for Disease Control (CDC) and Prevention d'Atlanta (homologue de Santé Publique France), comme une surveillance fondée sur une automatisation de l'enregistrement des données, permettant la mise à disposition pour le suivi et l'analyse épidémiologique en temps réel ou presque réel [32]. Créé en 1984, le réseau Sentinelles est composé de 1314 médecins généralistes et 116 pédiatres libéraux volontaires répartis sur le territoire métropolitain français. Il a pour mission de :

— construire de grandes bases de données en médecine générale et en pédiatrie, à des fins de veille sanitaire et de recherche,

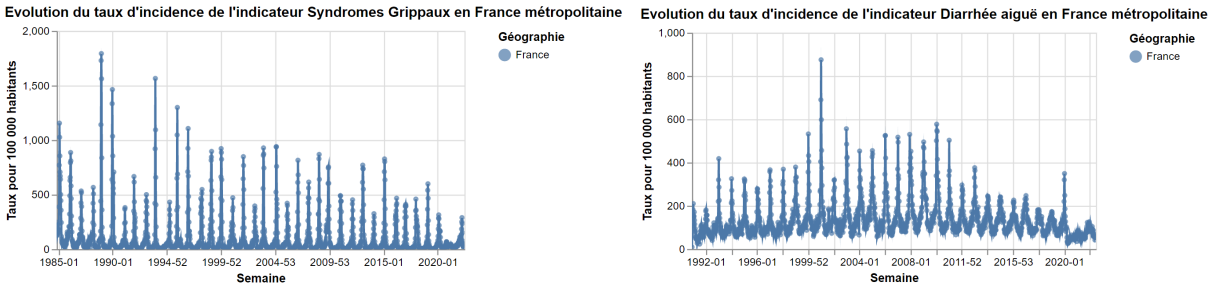


FIGURE 2.1 – Surveillance de la grippe et de la gastro-entérite par le Réseau Sentinelles

— développer des outils de détection et de prévision épidémique, et
— mettre en place des études cliniques et épidémiologiques.

Chaque semaine, les médecins Sentinelles saisissent le nombre de cas vus en consultations afin d'obtenir une surveillance continue sur 10 indicateurs de santé. Sont ainsi surveillées 9 maladies infectieuses :
— la coqueluche,
— la diarrhée aiguë,
— la maladie de Lyme,
— les oreillons,
— depuis 2020, les Infections Respiratoires Aiguës (IRA) incluant la COVID-19, grippe et autres virus respiratoires, définis par l'apparition brutale de fièvre et de signes respiratoires,
— les syndromes grippaux, dont la définition fut modifiée fin 2020 avec la COVID-19, sa définition est aujourd'hui un sous-ensemble des cas IRA,
— l'urétrite masculine, remplacée en 2020 par les infections sexuellement transmissibles bactériennes masculines et féminines,
— la varicelle,
— et le zona,

et 1 indicateur non-infectieux : les actes suicidaires.
À partir des données transmises, une estimation du taux d'incidence hebdomadaire est calculée pour chaque indicateur. Les séries chronologiques régionales ou nationales sont publiques. Pour la grippe et la gastro-entérite (figure 2.1), des cartes de France dynamiques visualisant les propagations épidémiques, et des rapports de prévision y sont publiés chaque semaine. Cependant, ces rapports ont un délai de 1 à 3 semaines en raison du temps de traitement et d'agrégation des données. Ce décalage est problématique pour des prises de décision optimales au niveau de l'agence nationale de santé publique [33, 34]. Afin d'apporter une aide à la décision, il est donc nécessaire de développer des méthodes permettant d'obtenir des estimations en temps réel et des estimations à plus long terme des taux d'incidence.

Un modèle autorégressif AR(52) a été appliqué sur ces données dans deux de nos articles à des fins de comparaison avec des modèles tenant compte de données supplémentaires de grande dimension qui seront présentées par la suite.
Nous appellerons "prédiction en temps réel" la prédiction à 0 semaine de Y_t à partir de $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-52}$.
Le modèle à d semaine(s) s'écrit :

$$Y_{r,t+d} = \sum_{j=1}^{52} \phi_j Y_{r,t-j} + \epsilon_{r,t+d} \quad \text{pour } d \in 0,1,2,3$$

où $Y_{i,t}$ correspond à l'incidence au temps t pour la région r, $Y_{r,t-j}$ avec j allant de 1 à 52 correspond

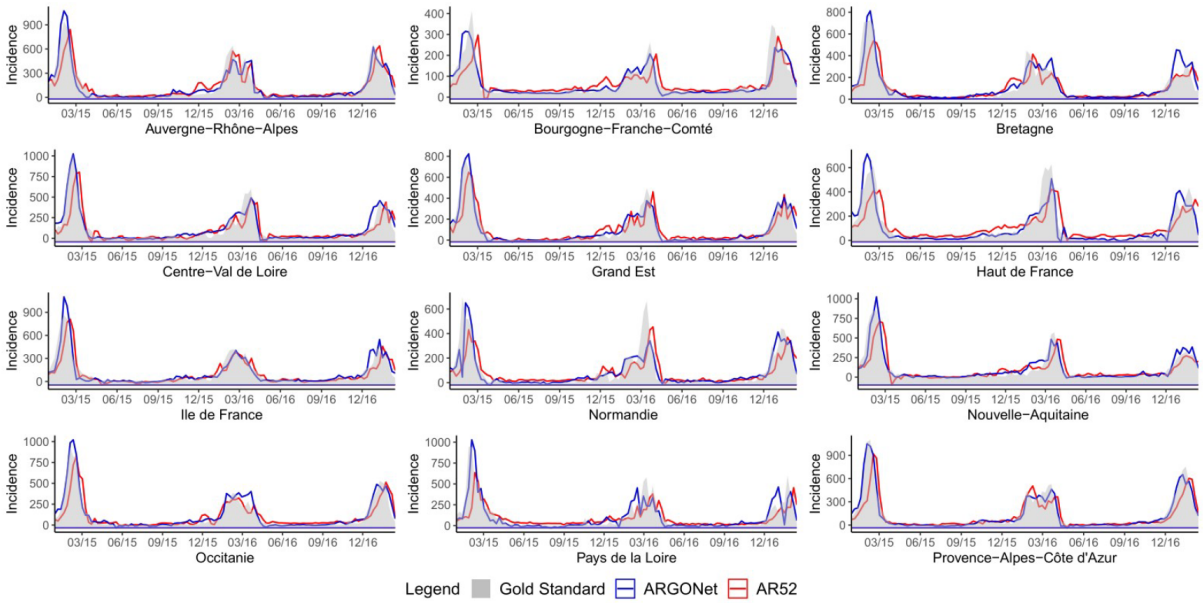


FIGURE 2.2 – Figure 7 de l'article "Grippe multi-sources" de thèse de Canelle Poirier [12]. Incidences grippales. Estimation en temps réel obtenue avec les modèles ARGONet et AR(52) de janvier 2015 à mars 2017.

aux 52 semaines précédentes, ϕ_j les 52 coefficients à estimer, et $\epsilon_{r,t+d}$ les résidus. Nous avons appliqué ce modèle pour chaque région, et au niveau national en n'indexant alors plus sur r. Nous avons utilisé une fenêtre glissante de 6 années pour l'échantillon d'apprentissage permettant un recalibrage dynamique du modèle pour chaque nouvelle information entrante.

Pour permettre la comparaison de notre modèle AR(52) avec les modèles qui seront présentés par la suite, les coefficients ϕ_j ont été estimés par l'algorithme LASSO (Least Absolute Shrinkage and Selection Operator) qui sera détaillé dans la partie "Modélisation de séries temporelles par régressions linéaires pénalisées".

La figure 2.2 montre en rouge les prédictions en temps réel par ces modèles versus en gris les données du réseau Sentinelles pour 12 régions de France. Les courbes rouges montrent des épidémies qui semblent un peu en retard et avec des pics légèrement sous-estimés, mais donnent déjà des résultats intéressants. Les comparaisons plus approfondies par différentes métriques, et les interprétations, seront détaillées plus loin.

La figure 2.3 montre en rouge les prédictions en temps réel, à une, deux et trois semaines par ces modèles versus en gris les données nationales d'incidences de gastro-entérite du réseau Sentinelles. Comme attendu, les courbes de prédiction s'éloignent de la courbe observée quand l'échéance de prédiction augmente.

2.2 Modélisation de séries temporelles par régressions pénalisées

Pour palier au délai de parution des estimations des réseaux de surveillance précédemment évoqué, des études se sont basées sur le Big Data et notamment sur les données du Web pour la prédiction en temps réel des épidémies. C'est le cas de Google, qui s'est rendu compte que certaines requêtes effectuées

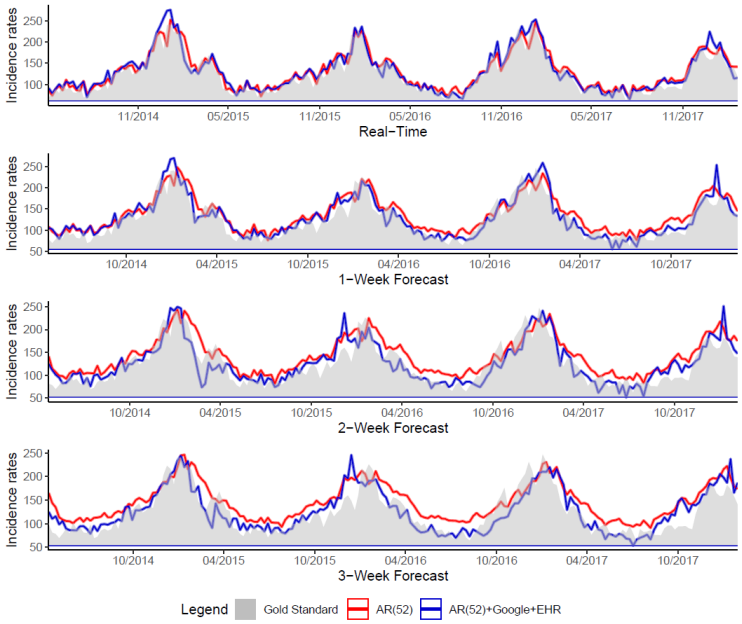


FIGURE 2.3 – Figure 1 de l'article "Gastro bi-sources" de thèse de Canelle Poirier [15]. Incidences de gastro-entérite. Prédictions jusqu'à trois semaines obtenues au niveau national avec le modèle utilisant uniquement l'historique et le modèle utilisant en plus les deux sources de données Google et EHR (eHOP).

par les internautes étaient fortement corrélées avec le taux d'incidence des épidémies de grippe.
En 2008, Google a alors proposé le modèle Google Flu Trends qui a permis de réaliser de bonnes prévisions pour les premières années. La série chronologique est comparée à environ 50 millions de requêtes jouées par les internautes que nous pouvons facilement imaginer variables dans le temps, ce qui permet d'identifier des "mots clé" les plus corrélés avec la série (45 requêtes ou "mots clé" les plus corrélées au signal sont conservées pour être utilisées lors de la modélisation des épidémies). Cette étape est réalisée au niveau national et régional. Cependant, l'épidémie de 2012-2013 a largement été surestimée, à cause de l'annonce d'une pandémie qui en réalité n'est pas apparue (figure 2.4). Cela a montré que le modèle n'était pas robuste car très sensible aux changements de comportement des internautes et à l'évolution de la performance du moteur de recherche. Google a donc mis fin à ce modèle.
De là, a été développé en 2015 le modèle ARGO (AutoRegression with Google search data) en utilisant conjointement les données de Sentinelles et les données de Google (article [35]), par l'équipe chez qui Canelle Poirier fera, quelques années plus tard, une mobilité de 6 mois durant sa thèse suivie d'un post-doctorat d'un an et demi.

2.2.1 Description du modèle ARGO

Il est composé de deux parties :

- Une première partie, appelée composante autorégressive, intégrant les données de Sentinelles américaines (CDC) avec N variables correspondant aux taux d'incidence des symptômes grippaux avec un décalage allant de 1 à N semaines.
- Une deuxième partie, appelée composante régressive, permettant d'intégrer les K requêtes de Google les plus corrélées au signal d'activité grippale.

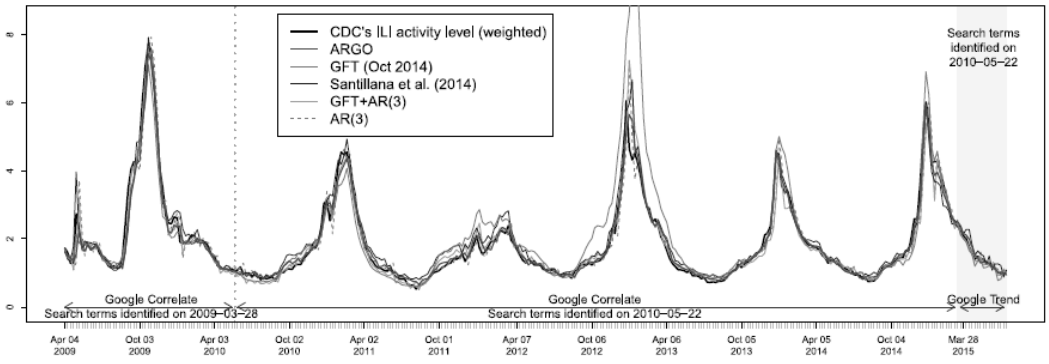


FIGURE 2.4 – Extrait de la figure 1 de l'article [35] des auteurs du modèle ARGO. Comparaisons de prédictions d'incidences grippales par différents modèles. Google Flu Trends (GFT (oct 2014)) surestime énormément l'épidémie 2012-2013.

$$Y_t = \beta_0 + \sum_{j=1}^{N} \phi_j Y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma^2)$$

où X_t peut être considéré comme des variables exogènes aux séries temporelles $\{Y_t\}$.
La partie autorégressive permet de prendre en compte l'historique des épidémies. La partie régressive, correspondant aux signaux de Google corrélés au signal de Sentinelles, a pour but de résoudre le problème de délai, mais aussi de variations inter-épidémiques. Ces données sont disponibles en temps réel. Tout comme Sentinelles, nous avons pour chaque semaine, un taux d'incidence correspondant au nombre de requêtes jouées par les internautes.
Les données de Sentinelles sont log-transformées et celles de Google sont logit-transformées pour les rendre plus normales. Le recalibrage dynamique décrit précédemment est également appliqué : pour chaque semaine à prédire, un nouveau modèle est crée à partir d'un jeu d'apprentissage correspondant à 2 ans d'historique pour les différents taux d'incidence. N est choisi fixé à 52 et K à 100. Le but est de trouver les paramètres β_0, $\phi = (\phi_1, \ldots, \phi_{52})$, et $\beta = (\beta_1, \ldots, \beta_{100})$ qui minimise

$$\sum_t \left(Y_t - \beta_0 - \sum_{j=1}^{52} \phi_j Y_{t-j} + \sum_{i=1}^{100} \beta_i X_{i,t} \right)^2 + \lambda_\phi \|\phi\|_1 + \eta_\phi \|\phi\|_2^2 + \lambda_\beta \|\beta\|_1 + \eta_\beta \|\beta\|_2^2$$

où $\lambda_\phi, \lambda_\beta, \eta_\phi$ et η_β sont des hyperparamètres qui seront optimisés par validation croisée.
Cette définition d'ARGO est celle où la sélection de variables est réalisée avec ElasticNet, mais en enlevant la norme 2, il peut également être programmé avec la méthode LASSO, nous reviendrons plus loin sur cette différence.
Dans l'article, le modèle ARGO est comparé à plusieurs modèles dont le modèle de Google Flu Trends, un modèle autorégressif d'ordre 3, le modèle de Google Flu Trends associé à un autorégressif d'ordre 3 et un modèle naïf. Pour toutes les épidémies, ont notamment été calculés la racine carré de l'erreur quadratique moyenne (RMSE) et le coefficient de corrélation de Pearson entre les prédictions et les observations. ARGO s'est révélé être meilleur avec de très bonnes performances (corrélation de l'ordre de 0.98). Cette efficacité de l'apport des données de Google en plus de l'historique de la série temporelle nous donnait des raisons de penser que le Big Data Hospitalier pourrait compléter ou remplacer les données du web.

2.2.2 Entrepôt de données biomédicales de l'HOPital de Rennes (eHOP)

L'Entrepôt de données biomédicales de l'HOPital de Rennes (eHOP) est une base de données massives qui centralise l'exhaustivité des informations des patients du CHU de Rennes (à présent près de 5 millions de patients, 130 millions de comptes rendus d'examens et d'hospitalisations, prescriptions médicamenteuses, données biologiques, etc... et 1,2 milliard de données structurées....). Le système eHOP comprend également des outils de recherche d'information et de fouille de données permettant d'exploiter le contenu de l'entrepôt dans des domaines variés : recherche clinique, étude épidémiologique, évaluation thérapeutique, pharmacovigilance, détection d'infection nosocomiale, analyse médico-économique, etc.

Dans une étape préliminaire, nous avons cherché à exploiter ces données pour montrer qu'elles pouvaient être pertinentes pour la surveillance syndromique. Pour cela, nous avons choisi comme cas d'usage la grippe. Nous avons réussi à montrer que les signaux extraits de l'entrepôt clinique, en lien avec le syndrome grippal, étaient très corrélés au signal estimé par les moyens de surveillance traditionnels [11]. Nous avons utilisé la régression périodique de Serfling [36] qui est actuellement utilisée par le réseau Sentinel pour identifier les périodes épidémiques de grippe [37].
Le principe consiste en la définition d'un seuil d'élagage correspondant au 85^e quantile, d'un intervalle de confiance à 95% unilatéral pour détecter le début (lorsque les données observées ont dépassé ce seuil pendant deux semaines consécutives) et la fin (lorsque les données observées sont inférieures au seuil pendant deux semaines consécutives) des épidémies de grippe. Nous avons ajusté le modèle de régression linéaire suivant pour toute la période d'étude :

$$Y_t = \beta_0 + \alpha.t + \beta_k.cos\left(\frac{2k\pi}{T}.t\right) + \gamma_k.sin\left(\frac{2k\pi}{T}.t\right) + \epsilon_t$$

où β_0 est une constante, α un terme linéaire, k un nombre harmonique, β_k et γ_k des termes périodiques.

La période T est égale à 52.18 semaines et k à 2. L'erreur résiduelle correspond au terme ϵ_t. Nous avons constaté que la requête la plus fortement corrélée avec les estimations du réseau Sentinelles était basée sur des rapports du service des urgences avec diagnostic final de grippe (figure 2.5), qui correspond à la $3^{\text{ème}}$ ligne "eHOP-Full text (emergency department)" avec un coefficient de corrélation de Pearson de 0.931.

Les requêtes sur données textuelles sont :
- Des requêtes en lien avec la grippe ou les syndromes grippaux avec les mots-clés :
  - "grippe"
  - "syndrome grippal"
  - "grippe" ou "syndrome grippal"
  - "grippe" ou "syndrome grippal" avec absence de "vaccin grippe" ou "vaccination"
  - "vaccin grippe"
  - "grippe" ou "syndrome grippal" aux urgences
- Des requêtes en lien avec les symptômes :
  - "fièvre" ou "pyrexie"
  - "courbatures" ou "douleurs musculaires"
  - fièvre" ou "pyrexie" ou "courbatures" ou "douleurs musculaires"
  - fièvre" ou "pyrexie" et "courbatures" ou "douleurs musculaires"
  - Des requêtes en lien avec les médicaments :
  - "Tamiflu"

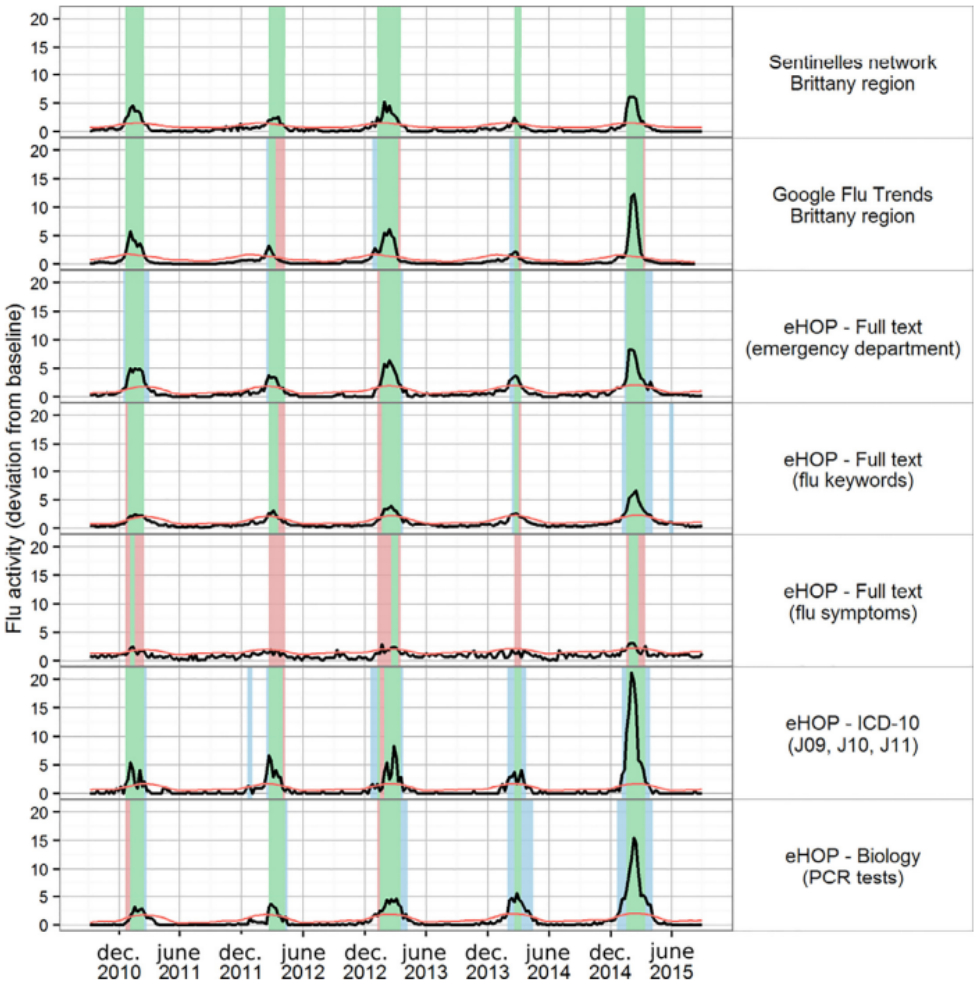


FIGURE 2.5 – Figure 2 du 1er article de thèse de Canelle Poirier [11]. Estimations hebdomadaires du syndrome grippal à partir des différentes sources de données et périodes d'épidémies détectées.

Dans un second temps, nous avons interrogé les données structurées grâce à deux terminologies.

— La terminologie CIM-10, qui est la classification internationale des maladies proposée par l'OMS. Elle permet de coder toutes les maladies et beaucoup de signes, symptômes, lésions ou encore traumatismes. Dans le cas de la grippe et du syndrome grippal, les codes sont : J09, J10 ou J11.
— Une terminologie locale utilisée par les laboratoires et permettant de retourner tous les résultats des tests PCR détectant la grippe.

Au total, en combinant les requêtes précédentes avec les mots clés "OU" et "ET", nous avons extrait 34 signaux de l'entrepôt de données eHOP. Dans certains modèles testés des articles *Grippe uni-source* [12] et *Grippe multi-sources* [13], ce nombre de 34 a été réduit à 10, ou même 3 signaux les plus corrélés.

Enfin pour l'application à la gastro-entérite de l'article *Gastro bi-sources* [15], tout comme pour la grippe, nous avons effectué des recherches sur les données textuelles grâce à des mots-clés en lien avec la gastro-entérite : ses symptômes, ses virus et ses traitements. De cette manière, nous avons obtenu 19 signaux

extraits de l'entrepôt eHOP. Nous avons également calculé, pour chaque niveau de prédiction, les 100 signaux les plus corrélés. Ces signaux ont été extraits d'une base de données contenant les séries temporelles construites à partir des données structurées de l'entrepôt de données eHOP. La corrélation a été calculée entre janvier 2008 et avril 2014. Au total, pour chaque niveau de prévisions, nous obtenons 119 variables explicatives. 19 variables au minimum sont communes à chaque niveau de prédiction.

2.2.3 Données publiques testées en variables exogènes

Données Google

La fréquence par semaine des 100 requêtes internet des français utilisateurs de Google les plus corrélées avec les données du réseau Sentinelles, ont été obtenues avec Google Correlate qui n'est plus disponible depuis décembre 2019 [38]. Les données Google ont été générées aux niveaux national et régional en fonction de l'échelle des incidences du réseau Sentinelles fournies en entrée de Google Correlate. Une fois les "mots-clés" les plus corrélés identifiés, nous avons utilisé le package R gtrendsR [39, 40] pour collecter le nombre de recherches quotidiennes françaises sur chaque "mot-clé".

Données Twitter

Les tweets géolocalisés ont été extraits, à l'échelle nationale pour la France, à partir de la base "Boston Children's Hospital Geotweet" avec les mots-clés suivants relatifs à la grippe (grippe, grippe', syndrome grippal, fièvre, toux, congestion, malade, faiblesse, courbatures, tamiflu, la crève). À partir de là, nous avons agrégé les tweets pour obtenir 11 variables Twitter de nombres de tweets.

Données météorologiques

Les données météorologiques spécifiques à la région proviennent du site Info Climat [1]. Il a été démontré dans plusieurs études que l'humidité est corrélée à la propagation de la grippe [41]. En l'absence de données d'humidité sur le site Info Climat, nous avons collecté les données quotidiennes de précipitations et température [42, 43] pour la plus grande ville de chaque région. Celles-ci peuvent être utilisées comme approximation de l'humidité puisqu'elles sont directement liées par la relation Clausius-Clapeyron [44]. Les moyennes hebdomadaires ont été calculées pour être utilisées dans les modèles.

2.2.4 Modèles construits pour les trois applications françaises (grippe et gastro-entérite)

Le modèle Elastic Net est un modèle de régression linéaire pénalisée (dont le principe est d'ajouter une contrainte sur les coefficients à estimer afin de pouvoir maîtriser l'amplitude de leurs valeurs), permettant de prendre en compte le grand nombre de variables explicatives et également la corrélation pouvant être présente entre ces variables [45]. Pour cela, elle combine les avantages de deux autres régressions linéaires pénalisées, les méthodes LASSO évoquée précédemment, et Ridge [46, 47].

Nous avons vu que pour LASSO, avec λ l'hyperparamètre à fixer afin de contrôler l'impact de la pénalité, le critère à minimiser était de la forme : $\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$.

Pour la régression Ridge, le critère est de la forme : $\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2$.

1. https ://www.infoclimat.fr

A la différence de la régression Ridge, la régression LASSO peut permettre d'effectuer une sélection de variables, en attribuant la valeur nulle à certains coefficients j, comme le montre la figure 2.6.

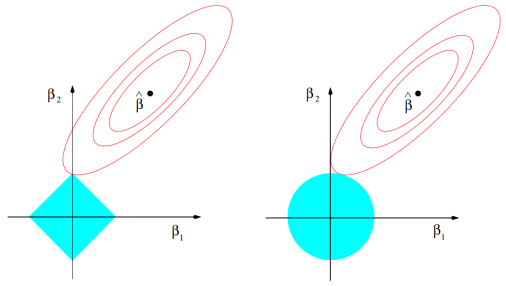


FIGURE 2.6 – FIGURE 3.11 du livre [48]. Contours des fonctions d'erreur de moindres carrés (ellipses rouges) et de contrainte (en bleu) des régressions Lasso (à gauche) et Ridge (à droite). Le losange correspond à la contrainte $|\beta_1|+|\beta_2| \leq t$ et le disque à $\beta_1^2 + \beta_2^2 \leq t$. Les deux méthodes trouvent la solution là où les contours elliptiques croisent la région de contrainte. Contrairement au disque, le losange a des coins, et donc la possibilité d'estimer des β_j à zéro (à gauche : $\beta_1 = 0$).

Pour notre cas, comportant une partie autorégressive et des covariables, avec ElasticNet et ses hyperparamètres λ et η, et β_0 fixé à 0, le critère à minimiser est :

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{d=1}^{D}\phi_d Y_{i-d} - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda_\phi \sum_{d=1}^{D}|\phi_d| + \eta_\phi \sum_{d=1}^{D}\phi_d^2 + \lambda_\beta \sum_{j=1}^{p}|\beta_j| + \eta_\beta \sum_{j=1}^{p}\beta_j^2.$$

Afin d'optimiser ces hyperparamètres, nous avons utilisé une validation croisée 10 blocs. Les modèles ont été réalisés grâce au package glmnet du logiciel R [49, 50], qui utilise la descente cyclique des coordonnées pour optimiser successivement la fonction objectif sur chaque paramètre, les autres étant fixés, et effectue des cycles répétés jusqu'à convergence.

Nom pour les différencier	Grippe uni-source (notre article [12])	Grippe multi-sources (notre article [13])	Gastro bi-sources (notre article [15])
Modèle	$Y_t = \sum_{i=1}^{D'}\phi_i Y_{t-i} + \sum_{d=0}^{D}\sum_{j=1}^{J}\beta_{j,d}X_{j,t-d} + \epsilon_t$	$Y_{r,t+d} = \sum_{i=1}^{D'}\phi_i Y_{r,t-i} + \sum_{j=1}^{J}\beta_j X_{j,r,t} + \sum_{l=1}^{L}\gamma_l V_{l,r,t} + \sum_{m=1}^{M}\delta_m W_{m,r,t} + \epsilon_{r,t+d}$	$Y_{t+d} = \sum_{i=1}^{D'}\phi_i Y_{t-i} + \sum_{j=1}^{J}\beta_j X_{j,t} + \sum_{k=1}^{K}\eta_k Z_{k,t} + \epsilon_{t+d}$
Régularisation	ElasticNet	LASSO	ElasticNet
Syndrôme	Grippe	Grippe	Gastro-entérite+Grippe
Prédictions	en temps réel	à d semaines ($d = (0, 1, 2)$)	à d semaines ($d = (0, 1, \ldots, 10)$)
Géographie	France + Bretagne	France + 12 régions r	France + Bretagne
Période	28/12/03 - 24/10/16	05/01/04 - 13/03/17	07/01/08 - 26/03/18
Partie AR	$\{D' = 52$ ou $2\}$; $\mathbf{Y_{t-i}}$	$D' = 52$; $\mathbf{Y_{t-i}}$	$D' = 52$; $\mathbf{Y_{t-i}}$
eHop	$\{D = 0; J = 34\}$ ou $\{D = 2; J = 3\}$; $\mathbf{X_j}$	$J' \in \{0, 1, \ldots, 10\}$; $\mathbf{X_{j,r,t}}$	$J = 119$; $\mathbf{X_{j,t}}$
Google	$\{D = 0; J = 100\}$ ou $\{D = 2; J = 3\}$; $\mathbf{X_j}$	$J'' \in \{0, 1, \ldots, 10\}$; $J = J' + J'' = 10$; $\mathbf{X_{j,r,t}}$	$K = 100$; $\mathbf{Z_{k,t}}$
Climatic		$L = 2$; $\mathbf{V_{l,r,t}}$	
Twitter		$M = 11$; $\mathbf{W_{m,r,t}}$	
Variante 1		Net : $Y_{r,t+d} = \sum_{l=1}^{2}\sum_{j=1}^{12}\phi_j Y_{j,t-l} + \sum_{j=1_{j\neq i}}^{12}\beta_j \hat{Y}_{j,t+d} + \epsilon_{r,t+d}$ où $\hat{Y}_{j,t+d}$ obtenus avec ARGO	
Variante 2		ARGONet : combinaison linéaire des estimations ARGO et Net	
Comparaisons	RF+SVM	AR(52)	AR(52)+RF+LSTM

TABLE 2.1 – Les modèles utilisés dans les 3 publications [12, 13, 15]

Le tableau 2.1 formalise l'écriture des modèles utilisés dans les 3 articles pour souligner leurs différences en termes de données modélisées et de données exogènes utilisées, en termes de délai de prédiction et de méthodes de régularisation, et cite les comparaisons de méthodes qui ont été réalisées.

Les résidus ϵ_i ont pu parfois être modifiés par $g(x_t, \epsilon_t)$ modélisés par un modèle ARIMA(p, d, q) pour respecter les hypothèses de bruit blanc et de stationnarité de la série.

2.2.5 Métriques d'évaluation des méthodes

Les quatre métriques que nous avons utilisées sont détaillées dans la table 2.2, nous avons :
- l'erreur quadratique moyenne (MSE),
- la corrélation (PCC),
- le décalage des pics ($\Delta(L)$), et
- la différence de hauteur des pics ($\Delta(H)$).

Pour illustrer en quoi ces indicateurs sont complémentaires, la figure 2.7 montre une corrélation parfaite entre la courbe rouge et la courbe noire, avec un $\Delta(L)$ égal à 0 (donc sans décalage de pics), mais un grand MSE et un grand $\Delta(H)$, alors que la courbe verte a une corrélation moindre, mais moins d'erreur quadratique (MSE plus faible donc meilleur), un $\Delta(H)$ et $\Delta(L)$ égaux à 0.

Métrique	Formules
MSE	$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
PCC	$\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(\hat{Y}_i-\hat{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \sum_{i=1}^{n}(\hat{Y}_i-\bar{\hat{Y}})^2}}$
$\Delta(L)$	$\text{argmax}_i\, Y_i - \text{argmax}_i\, \hat{Y}_i$
$\Delta(H)$	$\max Y_i - \max \hat{Y}_i$

Table 2.2 – Métriques d'évaluation des méthodes.

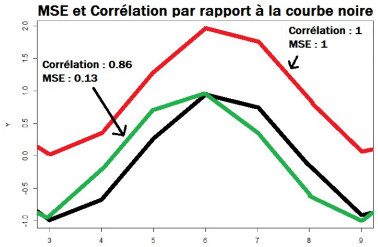


Figure 2.7 – Illustration de la différence entre les métriques

2.2.6 Visualisation par Heatmap des paramètres estimés

L'estimation des coefficients peut se visualiser avec une représentation thermique montrant les coefficients utilisés dans un modèle (figure 2.8).

Nous pouvons voir par exemple que ce modèle ARGO utilise toutes les 10 variables Google et hospitalières mais pas sur toute la période étudiée, la variable température et surtout une variable des données historiques Y_{t-1}.

FIGURE 2.8 – Figure 9 de l'article "Grippe multi-sources" de thèse de Canelle Poirier [13]. Visualisation des coefficients obtenus avec le modèle ARGO climatique hospitalo-web grippal pour la prédiction en temps réel avec chaque modèle pour une semaine donnée dans la région Nouvelle Aquitaine. Les 12 variables surlignées sont celles qui participent le plus au modèle : 7 variables de Google, 3 hospitalières, la température moyenne régionale et Y_{t-1}

2.3 Modélisation de séries temporelles par méthodes d'apprentissage statistique

2.3.1 Random Forest (RF)

Les forêts aléatoires ont été introduites en 2001 par Breiman [51] et sont, depuis, l'un des algorithmes les plus populaires en apprentissage supervisé [52]. La popularité vient du fait qu'ils sont rapides en temps de calcul et assez faciles à régler, et de la large gamme d'applications dans lesquelles ils sont connus pour bien fonctionner même en grande dimension, comme en chimie [53], en écologie [54], en reconnaissance d'objets 3D [55] et en prédiction de séries chronologiques [56, 57, 58] par exemple.

La méthode des forêts aléatoires est basée sur l'agrégation d'arbres de régression ou de discrimination, appelés aussi arbres de décision. Comme pour la régression, les arbres de décision sont utilisés pour la prédiction ou l'explication d'une variable cible Y, à partir d'un ensemble de variables explicatives X. Le principe des arbres est de diviser l'ensemble des données d'apprentissage successivement en sous-groupes. Les sous-groupes doivent être le plus homogène possible, les divisions se font grâce aux variables explicatives, qui, à chaque étape, discriminent au mieux la variable cible. Les sous-groupes intermédiaires de la variable Y sont appelés nœuds et les sous-groupes finaux sont appelés feuilles. Dans notre étude, nous avons utilisé des arbres de régression car nous cherchions à prévoir une variable quantitative, les taux d'incidence de grippe. Les sous-groupes ont été construits grâce aux variables issues du réseau Sentinelles et aux variables exogènes (eHOP et/ou Google). Afin d'obtenir des sous-groupes les plus homogènes possibles, il est nécessaire de choisir en priorité les variables qui vont diminuer la variance intra-classe. Les valeurs des nœuds et des feuilles vont correspondre à la moyenne des taux d'incidence de grippe composant les sous-groupes. Cependant les arbres souffrent d'une grande instabilité (fléau de la dimension, sensibilité à l'échantillonnage), c'est pour cette raison qu'il est nécessaire d'utiliser une méthode d'agrégation. La forêt aléatoire (RF) est donc basée sur la construction de plusieurs arbres de décision à l'aide de la technique d'agrégation bootstrap (connue sous le nom de bagging, développé en 1996 également par Breiman). Les taux d'incidence sont obtenus avec :

$$Y_T = \frac{1}{B} \sum_{b=1}^{B} \hat{Y}_b$$

où Y_T le taux d'incidence au temps $T = [t; t+1; t+2; t+3]$ (pour les différentes échelles de prédiction) ;$\hat{Y}_b$ le taux d'incidence estimé obtenu avec l'arbre de décision b. Chaque arbre se distingue par le sous-échantillon de données sur lequel il est entraîné. En effet, pour chaque arbre, nous construisions un échantillon bootstrap en tirant aléatoirement avec remise, un nombre N d'observations identique à celui des données d'origine. De plus, une 2ème partie aléatoire est ajoutée, en ne choisissant que m variables explicatives parmi toutes les variables disponibles. Cet hyperparamètre m a été choisi par validation croisée 10 blocs.
Nous avons utilisé le package R randomForest [59] pour créer nos modèles RF.
La figure 2.9 montre les erreurs et la corrélation pour la grippe et la gastro-entérite aux niveaux national et régional, pour les prévisions jusqu'à 10 semaines (notre article [15]).

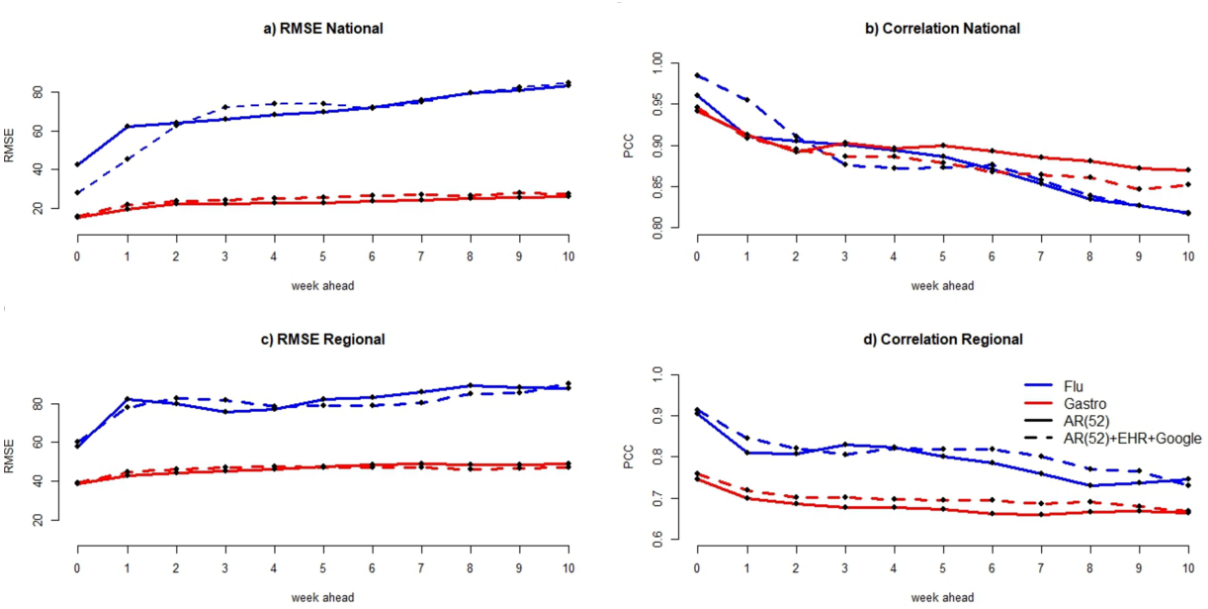


FIGURE 2.9 – Figure 9 de l'article "Gastro bi-sources" [15]. Prévisions jusqu'à 10 semaines avec le modèle Random Forest (RF) pour la grippe et la gastro-entérite. La ligne pleine correspond aux résultats obtenus avec le modèle Elastic Net en utilisant uniquement des données historiques. La ligne pointillée correspond aux résultats obtenus avec le modèle RF en utilisant des données historiques et des données Google et eHOP. Sont en rouge les résultats pour la gastro-entérite et en bleu la grippe. a) Valeurs d'erreur obtenues au niveau national. b) Valeurs de corrélation obtenues au niveau national. c) Valeurs d'erreur obtenues au niveau régional. d) Valeurs de corrélation obtenues au niveau régional.

Pour la gastro-entérite, au niveau national, en termes d'erreurs, les valeurs les plus faibles sont obtenues en utilisant uniquement des données historiques (AR(52)), alors qu'au niveau régional, c'est le cas seulement pour les 4 premières semaines, ensuite les résultats les plus précis sont obtenus à l'aide de données historiques et de sources de données, Google et eHOP.
En termes de corrélation, au niveau régional, les données exogènes apportent au modèle, alors qu'au niveau national les valeurs les plus faibles sont obtenues en utilisant uniquement des données historiques.
Pour la grippe, les courbes d'erreurs se croisent en 2 points (ce qui se vérifie aussi sur les corrélations) : au niveau national nous voyons un apport des données exogènes jusqu'à 3 semaines, entre trois à cinq semaines, utiliser les données historiques seules dans nos Random Forest est meilleur, pour les prévisions à long terme, les résultats sont semblables pour les deux modèles. Au niveau régional, globalement, les valeurs les plus élevées sont obtenues en utilisant des données historiques et des données Google et eHOP avec une petite entorse entre deux et 4 semaines.
Bien évidemment entre grippe et gastro-entérite, les erreurs ne sont pas comparables puisque dépendantes de l'ordre de grandeur des incidences. Par contre la comparaison des corrélations est intéressante. Au niveau national, il y a une chute très importante de la corrélation pour la grippe jusqu'à 0.82 pour les prédictions à 10 semaines, alors que la gastro est plus stable avec une corrélation plus faible que la grippe pour les premières semaines puis une inversion sur les dernières semaines.
Au niveau régional, l'écart entre la grippe et la gastro est important, et la corrélation pour la gastro ne se situe qu'autour de 0.7.
Nous confirmerons ou infirmerons ces tendances par la méthode linéaire "modèle Gastro bi-sources" décrite précédemment dans le tableau 2.1, pour lequel les résultats correspondants seront présentés dans la figure 2.17.

Toutefois, lorsqu'il s'agit de séries chronologiques, les forêts aléatoires n'intègrent pas la structure dépendante du temps, supposant implicitement que les observations sont indépendantes. Plusieurs variantes des forêts aléatoires ont été récemment conçues pour séries chronologiques [60]. Notre méthode de fenêtres glissantes permet de limiter la baisse de performance, mais ces nouveaux algorithmes permettraient peut-être d'en gagner.

2.3.2 Support Vector Machine (SVM)

Les Support Vector Machines (SVM)[61], qu'on appelle souvent en français Séparateur à Vaste Marge (pour garder l'acronyme) sont des méthodes basées sur des algorithmes d'apprentissage, qui consistent à ramener le problème de la discrimination à celui de la recherche d'un hyperplan optimal permettant de séparer l'espace dans lequel les données prennent leurs valeurs. Les SVM, développées au départ pour la classification supervisée binaire, ont rapidement vu naître des extensions, comme les SVR (Support Vector Regression) qui permettent la prévision d'une variable quantitative.

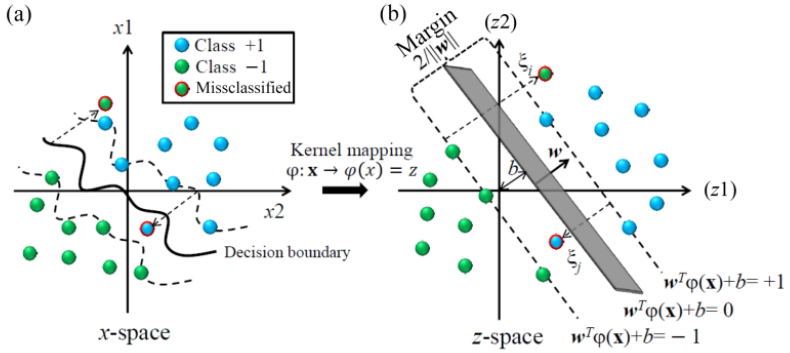


FIGURE 2.10 – Figure 1 de l'article [62]. Illustration de la méthode SVM avec un noyau non linéaire : (a) le problème complexe de classification supervisée binaire dans l'espace d'entrée, et (b) la notion de vaste marge.

Pour décrire brièvement le principe des SVM, la figure 2.10 montre le principe des vecteurs support et de la vaste marge, où l'on cherche :

$$\underset{w,b,\xi}{\operatorname{argmin}}\Big(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i\Big)$$

sous les contraintes
$\begin{cases} y_i(< w, x_i > +b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n; \\ \xi_i \geq 0, \quad \forall i = 1, \ldots, n; \end{cases}$
avec C un hyperparamètre qui sera à calibrer, correspondant à un paramètre de coût (appelé constante de tolérance). Plus C augmente plus l'ajustement aux données augmente.

Puisque nos Y_i sont des incidences, la variable à modéliser n'est pas binaire mais continue. Ce sera donc les SVR que nous utiliserons, même si dans nos applications, nous garderons par abus de langage le terme SVM plus parlant.

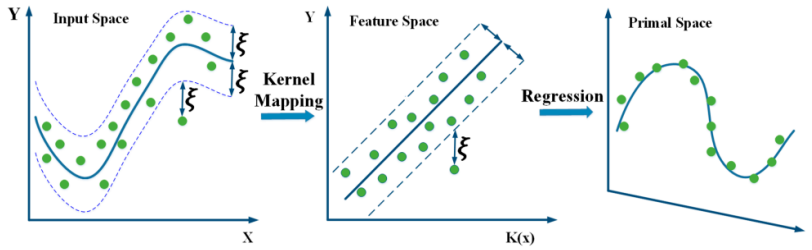


FIGURE 2.11 – Figure 2 de l'article [63]. Le principe de la régression à vecteur support (SVR).

Le principe est très similaire, la figure 2.11 montre la vaste marge construite autour des observations avec la tolérance ou plutôt pénalité ξ d'être en dehors de la marge, tout comme était tolérés des mal classés en SVM.

La figure 2.12 montre les résultats de notre application des SVM aux données d'incidences grippales [12]. La comparaison des données sentinelles en vert et des prédictions par SVM en pointillé bleu montre de belles performances, que nous détaillerons un peu plus dans la partie 2.4.

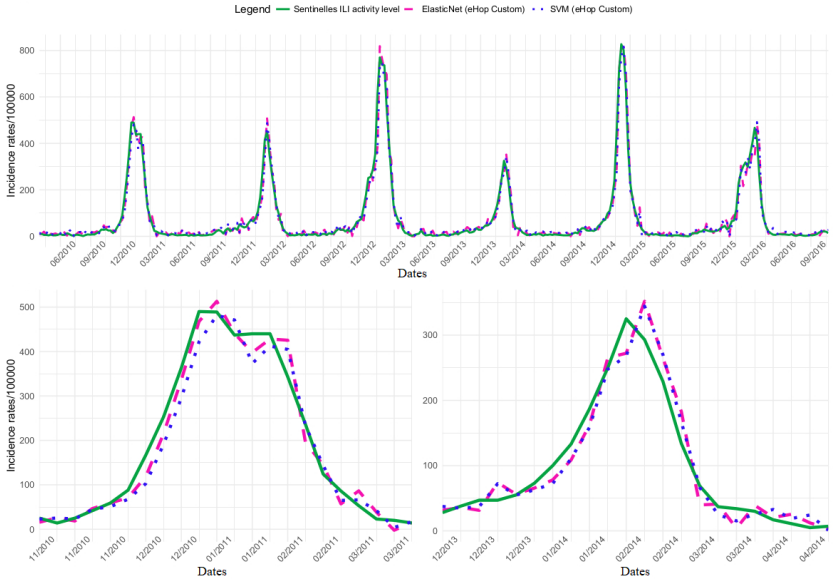


FIGURE 2.12 – Figure 2 de l'article "Grippe uni-sources" [12]. Estimations rétrospectives de l'activité nationale des syndromes grippaux (SG) obtenues à l'aide du modèle ElasticNet avec 3 des 34 variables hospitalières en plus des 52 variables historiques (eHOP Custom) en rose, et la méthode SVM en bleu, comparés aux données observées par le réseau Sentinelles en vert. La série entière et les épidémies de 2010-2011 et 2013-2014 sont présentées.

2.3.3 Réseau de neurones Long short-term memory (LSTM)

Cette partie sera très courte, puisqu'elle n'est pas au cœur de notre étude dans l'article [15], et reprend les grandes lignes du livre [64]. Pour situer le contexte, nous dévions maintenant vers l'apprentissage profond

(deep learning), qui peut être défini comme tout algorithme qui repose sur l'utilisation de neurones artificiels ou de couches. Pour lister rapidement les principaux :

— les réseaux de neurones artificiels, qui sont des imitations simples des fonctions d'un neurone dans le cerveau humain, et permettent de construire facilement des modèles très complexes et non linéaires,
— les convolutifs, dont le rôle est de pouvoir extraire de l'information cachée dans les images, par exemple reconnaître les caractéristiques d'un objet,
— les GAN, qui utilisent souvent des réseaux de neurones convolutifs mais sont plus puissants,
— les neurones récurrents, qui sont capables de garder en mémoire des informations, et donc d'un niveau de complexité plus grand encore, adaptés notamment aux prédictions de séries temporelles.

Le défi de la préservation de l'information à long terme et du saut d'entrée à court terme dans les modèles de variables latentes existe depuis longtemps. L'une des premières approches visant à relever ce défi a été le long short-term memory (LSTM) [65], une extension des réseaux de neurones récurrents. Le LSTM introduit une cellule mémoire qui a la même forme que l'état caché, conçue pour enregistrer des informations supplémentaires. Pour contrôler la cellule mémoire, nous avons besoin d'un certain nombre de portes.

— la porte de sortie, nécessaire pour lire les entrées de la cellule.
— la porte d'entrée, nécessaire pour décider quand lire les données dans la cellule.
— la porte d'oubli, nécessaire pour réinitialiser le contenu de la cellule.

La motivation d'une telle conception est d'être capable de décider quand se souvenir et quand ignorer les entrées dans l'état caché via un mécanisme dédié. Les données qui alimentent les portes LSTM sont l'entrée au temps t et l'état caché au temps $t-1$, comme illustré à la figure 2.13. Elles sont traitées par trois couches entièrement connectées avec une fonction d'activation sigmoïde pour calculer les valeurs des portes d'entrée, d'oubli et de sortie. Par conséquent, les valeurs des trois portes se situent dans l'intervalle [0,1].

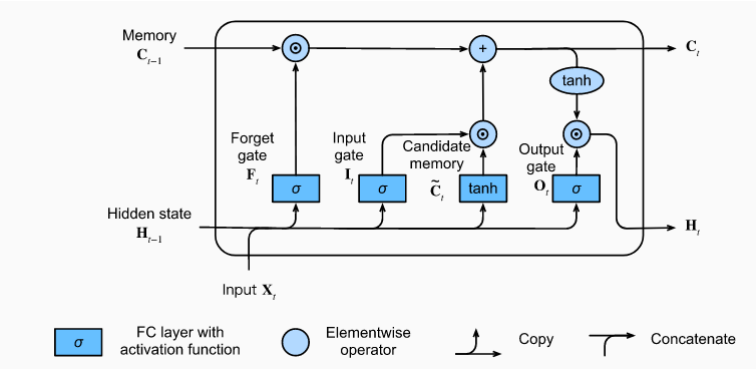


FIGURE 2.13 – Fig. 10.2.4 du livre [64]. Comprendre les différentes étapes du modèle LSTM.

Notons h le nombre de couches cachées, l'entrée est $X_t \in \mathbb{R}^{n\times p}$ et l'état caché à $t-1$ est $\mathbf{H}_{t-1} \in \mathbb{R}^{n\times h}$. En conséquence, les portes au pas de temps sont définies comme suit : la porte d'entrée est $\mathbf{I}_t \in \mathbb{R}^{n\times h}$, la porte d'oubli est $\mathbf{F}_t \in \mathbb{R}^{n\times h}$, et la porte de sortie est $\mathbf{O}_t \in \mathbb{R}^{n\times h}$. Elles sont calculées avec :

$$\mathbf{I}_t = \sigma(\mathbf{X}_t\mathbf{W}_{xi} + \mathbf{H}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i),$$

$$\mathbf{F}_t = \sigma(\mathbf{X}_t\mathbf{W}_{xf} + \mathbf{H}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f),$$

$$\mathbf{O}_t = \sigma(\mathbf{X}_t\mathbf{W}_{xo} + \mathbf{H}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o),$$

où $\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xo} \in \mathbb{R}^{p \times h}$ et $\mathbf{W}_{hi}, \mathbf{W}_{hf}, \mathbf{W}_{ho} \in \mathbb{R}^{h \times h}$ sont respectivement les poids de connexion des 3 couches à l'entrée x_t et à l'état court terme, et $\mathbf{b}_i, \mathbf{f}_i et \mathbf{b}_o \in \mathbb{R}^{1 \times h}$ sont les paramètres de biais. σ est la fonction d'activation logistique conduisant à des sorties qui varient entre 0 et 1 pour contrôler les portes. Regardons ensuite la cellule mémoire candidate $\widetilde{\mathbf{C}}_t \in \mathbb{R}^{n \times h}$. Son calcul est similaire à celui des trois portes décrites ci-dessus, mais en utilisant comme fonction d'activation une fonction *tanh* avec une plage de valeurs comprises dans [-1,1]. La fonction tangente hyperbolique est souvent choisie pour représenter un phénomène de transition progressive, « douce », entre deux états. Cela conduit à l'équation suivante en t :

$$\widetilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t \mathbf{W}_{xc} + \mathbf{H}_{t-1} \mathbf{W}_{hc} + \mathbf{b}_c),$$

où $\mathbf{W}_{xc} \in \mathbb{R}^{p \times h}$ et $\mathbf{W}_{hc} \in \mathbb{R}^{h \times h}$ sont respectivement les poids de connexion à l'entrée x_t et à l'état court terme, et $\mathbf{b}_c \in \mathbb{R}^{1 \times h}$ le paramètre de biais.

Dans les LSTM, nous disposons d'un mécanisme qui régit l'entrée et l'oubli : la porte d'entrée $\mathbf{I}_t$ régit le degré de prise en compte des nouvelles données via $\widetilde{\mathbf{C}}_t$ et la porte d'oubli traite de la quantité de contenu de l'ancienne cellule mémoire $\mathbf{C}_{t-1} \in \mathbb{R}^{n \times h}$ que nous conservons. En utilisant la même astuce de multiplication ponctuelle que précédemment, nous obtenons l'équation de mise à jour suivante :

$$\mathbf{C}_t = \mathbf{C}_t \otimes \mathbf{C}_{t-1} + \mathbf{I}_t \otimes \widetilde{\mathbf{C}}_t$$

Si la porte d'oubli est toujours approximativement égale à 1 et la porte d'entrée est toujours approximativement égale à 0, les cellules de mémoire passées $\mathbf{C}_{t-1}$ seront sauvegardées au fil du temps et transmises à t. Cette conception est introduite pour mieux capturer les dépendances à long terme dans les séquences. Enfin, il reste à définir comment calculer l'état caché $\mathbf{H}_t \in \mathbb{R}^{n \times h}$, c'est là que la porte de sortie entre en jeu. Dans les LSTM, il s'agit simplement d'une nouvelle utilisation de la fonction tangente hyperbolique. Cela garantit que les valeurs de $\mathbf{H}_t$ sont toujours dans l'intervalle [0,1].

$$\mathbf{H}_t = \mathbf{O}_t \otimes \tanh(\mathbf{C}_t)$$

Lorsque la porte de sortie se rapproche de 1, nous transmettons effectivement toutes les informations de la mémoire au prédicteur, tandis que pour la porte de sortie proche de 0, nous conservons toutes les informations uniquement dans la cellule de mémoire et n'effectuons aucun traitement supplémentaire.

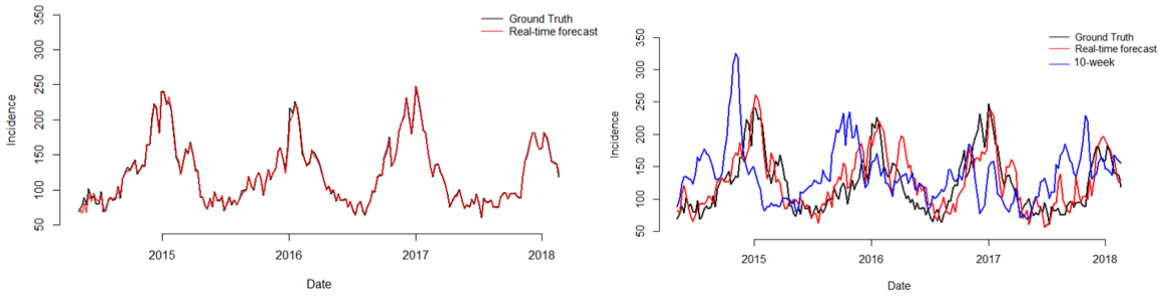


FIGURE 2.14 – Figures non publiées mais montrées aux 3 reviewers de l'article "Gastro bi-sources" de thèse de Canelle Poirier. Résultats obtenus au niveau national avec le modèle LSTM utilisant seulement les données historiques pour prédire uniquement l'incidence à une semaine à gauche ; et pour prédire à une et 10 semaines à droite.

La figure 2.14 montre une très belle performance de prédiction en temps réel, meilleure qu'un AR(52) classique. L'intérêt des LSTM est de pouvoir travailler plusieurs sorties à la fois, par exemple sur le graphique de droite la prédiction en temps réel et la prédiction à 10 semaines. Ce qui alors fait perdre en performance sur les deux types de prédiction. Ce modèle doit donc encore être affiné en rajoutant des couches. Nous pensons que ces résultats sont vraiment prometteurs. Cependant, l'approche d'apprentissage profond prend beaucoup de temps et nous perdons l'avantage de pouvoir détecter les variables les plus importantes pour prédire la gastro-entérite. De plus, nous pensons qu'il pourrait s'agir d'un sujet de recherche entier pour adapter le meilleur réseau neuronal à la prévision à plus long terme, car il serait nécessaire de développer un autre réseau neuronal si nous voulons inclure des sources de données externes telles que les données Google ou Twitter.

2.4 Résultats des comparaisons de modèles pour la prédiction de la grippe et de la gastro-entérite

Après avoir montré que les signaux extraits des entrepôts de données biomédicaux étaient corrélés aux signaux des réseaux de surveillance traditionnels (décrit dans la partie 2.2.2 et publié dans l'article [11]), l'objectif de notre premier article de modélisation [12] était de montrer l'intérêt des données massives hospitalières (eHOP) par rapport aux données du web (Google), et d'évaluer la performance de différents modèles statistiques pour la prévision des épidémies de grippe en temps réel en comparant les forêts aléatoires (RF), les Séparateurs à Vaste Marge (SVM) et le modèle de régression linéaire Elastic Net ("Modèle Grippe uni-source" du tableau 2.1).
Les meilleures estimations au niveau national et régional ont été obtenues grâce aux données hospitalières et au modèle SVM, même si Elastic Net donne des résultats similaires. Le coefficient de corrélation obtenu était égal à 0.98 et l'erreur quadratique moyenne était égale à 866 (tableau 2.3). Au niveau de la région Bretagne, le meilleur coefficient de corrélation était égal à 0.923 et l'erreur quadratique moyenne était égale à 2364.
Les résultats du tableau se visualisait également sur la figure 2.12 où pour l'épidémie 2010/2011, le pic est légèrement sous-estimé par SVM ($\Delta H = -8$) et légèrement sur-estimé par ElasticNet ($\Delta H = 23$), et tous les deux légèrement en retard ($\Delta L = 1$).

Cette étude nous a permis de montrer que les données massives hospitalières permettaient d'obtenir dans la plupart des cas, des prévisions en temps réel plus précises que les données du web. De plus, nous avons pu voir que deux modèles avaient des performances comparables, le modèle SVM et le modèle de régression linéaire pénalisé Elastic Net. Il serait intérressant de savoir dans quelle mesure les estimations pourraient être améliorées si nous combinions ces deux sources de données ou les résultats de différents modèles statistiques, mais aussi de prédire les épidémies à plus long terme et à une échelle de résolution plus fine. L'équipe de Fred S.Lu et al [66] a repris le modèle ARGO de Yang et al. [35] présenté dans la partie 2.2.1 en l'adaptant afin d'estimer en temps réel les épidémies de grippe au niveau de chaque état des États-Unis, à partir des données de Google, des données historiques du CDC, et des données de dossiers patients électroniques provenant d'une compagnie américaine Athenahealth. Dans cette même étude, ces chercheurs ont développé deux autres modèles, le modèle Net et un modèle combiné ARGONet. Le modèle Net est un modèle se basant sur la corrélation des épidémies entre les différents états et le modèle ARGONet est un modèle utilisant les estimations des deux autres approches ARGO et Net (écritures pour notre modèle "Grippe multi-sources" dans le tableau 2.1).
L'objectif de notre deuxième article de modélisation [13] est d'étendre aux régions françaises, les approches ARGO, Net et ARGONet développées aux États-Unis. Pour cela, nous avons utilisé les données de Google, les données massives hospitalières provenant de l'entrepôt de données biomédicales eHOP et

a. National	2010-2011				2013-2014				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	$\lvert\Delta H\rvert$	ΔL	$\lvert\Delta L\rvert$
Ehop Custom																
RF	0.95	4119	50	2	0.91	2212	75	1	0.947	2292	0.9	6916	−22	72	1.33	1.33
SVM	0.97	1932	−8	1	**0.95**	**996**	19	1	**0.98**	**866**	0.96	2716	6	19	0.83	0.83
Elastic + Arima	**0.98**	**1222**	23	1	0.95	1145	27	1	0.98	872	**0.96**	**2664**	26	30	0.66	0.66
Google Complete																
RF	0.95	2743	23	0	0.93	4931	84	1	0.963	1706	0.94	5764	−14	70	0.66	0.66
SVM	0.95	2671	12	3	0.92	**1564**	56	1	0.974	1192	**0.96**	**2805**	37	49	0.8	0.8
Elastic + Arima	**0.96**	**2153**	6	1	**0.95**	2511	85	1	0.978	1057	0.96	2967	39	38	0.66	0.66

b. Regional	2010-2011				2013-2014				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	$\lvert\Delta H\rvert$	ΔL	$\lvert\Delta L\rvert$
Ehop Custom																
RF	0.91	5796	−29	−1	0.65	4577	−4	1	0.911	2777	0.84	6929	−40	42	−0.2	1.5
SVM	0.92	5502	−53	1	**0.76**	**2477**	−28	1	**0.923**	**2364**	**0.86**	**6050**	−60	60	0.3	1
Elastic + Arima	**0.92**	**4689**	−28	0	0.71	2855	−27	1	0.918	2451	0.84	5999	−32	38	0.3	0.7
Google Complete																
RF	**0.92**	**4650**	−80	0	0.63	5955	−9	2	**0.912**	**2736**	**0.83**	**7122**	−62	62	0.7	1.7
SVM	0.84	8664	−97	1	0.56	4735	−43	−1	0.890	3348	0.70	9137	−52	59	0	1.67
Elastic + Arima	0.89	6455	−63	1	**0.78**	**2113**	−26	1	0.903	2967	0.79	7239	−31	32	0.7	1

TABLE 2.3 – Extrait de l'annexe 8 de l'article "Grippe uni-source" [12]. Mesures de concordance pour la période 2010-2011 (période épidémique de grippe pour laquelle les estimations ont été les meilleures avec tous les modèles) et 2013-2014 (période épidémique avec les pires estimations). PCC (corrélation) et MSE (erreur quadratique moyenne) pour la série temporelle de validation entière (Global) et les valeurs moyennes des périodes épidémiques (Means) de tous les indicateurs pour chaque modèle pendant les périodes épidémiques. En gras, les meilleurs résultats pour chaque jeu de données. a. Données pour toute la France. b. Données pour la région Bretagne.

les données historiques du réseau Sentinelles. Nous avons également étudié l'ajout d'autres sources de données comme les données climatiques et les données de Twitter. Enfin, nous avons dans cet article réalisé des prévisions jusqu'à 2 semaines.

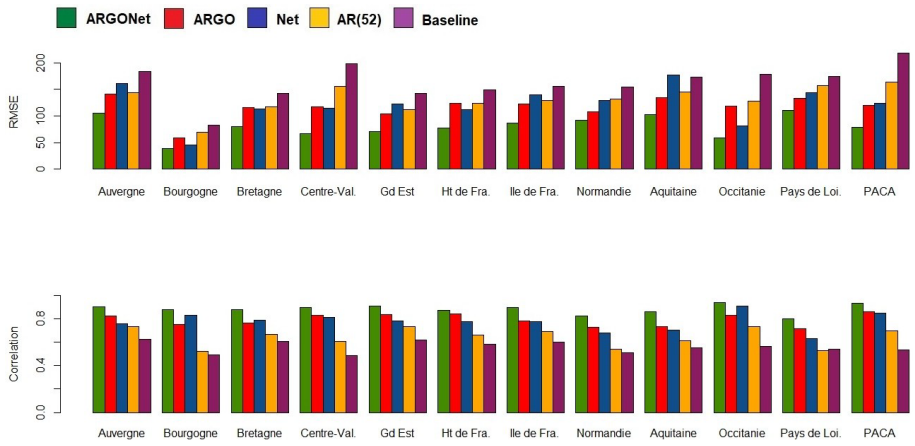


FIGURE 2.15 – Figure 10 de l'article "Grippe multi-sources" [13]. Visualisation des corrélations et erreurs obtenus pour les estimations à une semaine avec chaque modèle.

La figure 2.15 montre que les meilleures corrélations et erreurs quadratiques moyennes sont principalement obtenus avec ARGONet, mais ARGO donne également de bons résultats. En regardant par exemple la corrélation la plus faible de AR(52) obtenue pour la Bourgogne, la figure 2.2 permet de l'expliquer par les prédictions d'épidémies très décalées pour les épidémies 2014/15 et 2015/16, alors que la performance de ARGONet est très correcte.
La figure 2.16 montre que l'écart de performance d'ARGONet s'accentue quand on prédit à plus long

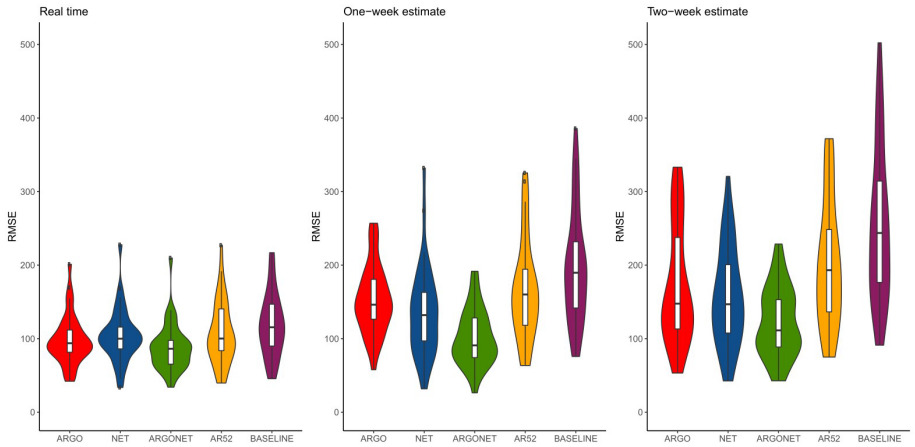


FIGURE 2.16 – Figure 14 de l'article "Grippe multi-sources" [13]. Visualisation des erreurs obtenus pour les estimations en temps réel, à une et deux semaines avec chaque modèle en région Nouvelle-Aquitaine.

terme, l'exemple de la Nouvelle-Aquitaine n'est pas isolé.
Dans cet article nous avons montré que, pour toutes les régions de France, le modèle ARGONet était le modèle le plus robuste pour prévoir les taux d'incidence grippaux jusqu'à 2 semaines et que l'utilisation de toutes les sources de données externes permettait d'améliorer les prévisions et plus particulièrement les prévisions à plus long terme. Pour conclure, la méthode d'ensemble ARGONet développée aux États-Unis pourrait être utilisée afin de compléter les méthodes de surveillance traditionnelles.

Nous avons ensuite souhaité savoir si les modèles développés précédemment, pouvaient être généralisés à d'autres maladies telles que la gastro-entérite, qui est également un enjeu de santé publique majeur à travers le monde [67]. Même s'il s'agit généralement d'une maladie bénigne, elle a une morbidité et un poids économique élevés. La diarrhée est l'une des principales causes de mortalité chez les jeunes enfants. Chaque année, à travers le monde, la gastro entérite est responsable d'environ 2.5 millions de décès chez les enfants de moins de 5 ans [68]. En France, chaque année, il y a plus de 21 millions de cas déclarés [69]. Tout comme la grippe, pendant les périodes épidémiques, il y a une forte augmentation de visites chez les médecins généralistes et au service des urgences, ce qui pose problème pour l'organisation des systèmes de santé. En France, c'est également le réseau Sentinelles qui est en charge de la surveillance des cas de diarrhée aiguë. Les méthodes d'estimation des taux d'incidence sont identiques à celles de la grippe, le problème de délai dû au traitement et à l'agrégation des données est donc également présent. A l'heure actuelle, des études se sont intéressées aux caractéristiques des épidémies de gastro entérite, comme le virus en circulation, l'impact du vaccin, les changements climatiques, mais peu d'études se sont intéressées à la prévision.
Notre article [15] présente deux approches d'apprentissage automatique (le modèle linéaire "Modèle Gastro bi-sources" (tableau 2.1 et le modèle non linéaire Random Forest) qui produisent des estimations en temps réel, des prévisions à court terme, et les prévisions à long terme de l'activité des gastro-entérites aiguës à deux échelles spatiales différentes en France (nationale et régionale). Les deux approches tirent parti de sources de données disparates, notamment : les données de Google, les données hospitalières, et l'activité historique de la maladie. Pour les prévisions à long terme, la Random Forest était plus performante (figure 2.9) que le modèle linéaire "Modèle Gastro" surtout sur la gastro-entérite qui voit sa corrélation chuter à 0.55 pour les prévisions à 10 semaines. La figure 2.17 montre parallèlement que les données exogènes contribuent alors à améliorer la surveillance de la gastro-entérite. Cet apport des données exogènes dans les comparaisons de prédiction jusqu'à 3 semaines se voyait déjà à l'oeil nu sur la

figure 2.3 où la dernière courbe rouge sur-estime énormément les incidences.
Nous avons également comparé les performances de prédictions gastro à celles de la grippe. Pour la gastro, le pouvoir prédictif des données historiques est très important en raison de la forte dynamique saisonnière de cette maladie, que l'on peut voir sur le graphique des auto-corrélations (figure 2.18).

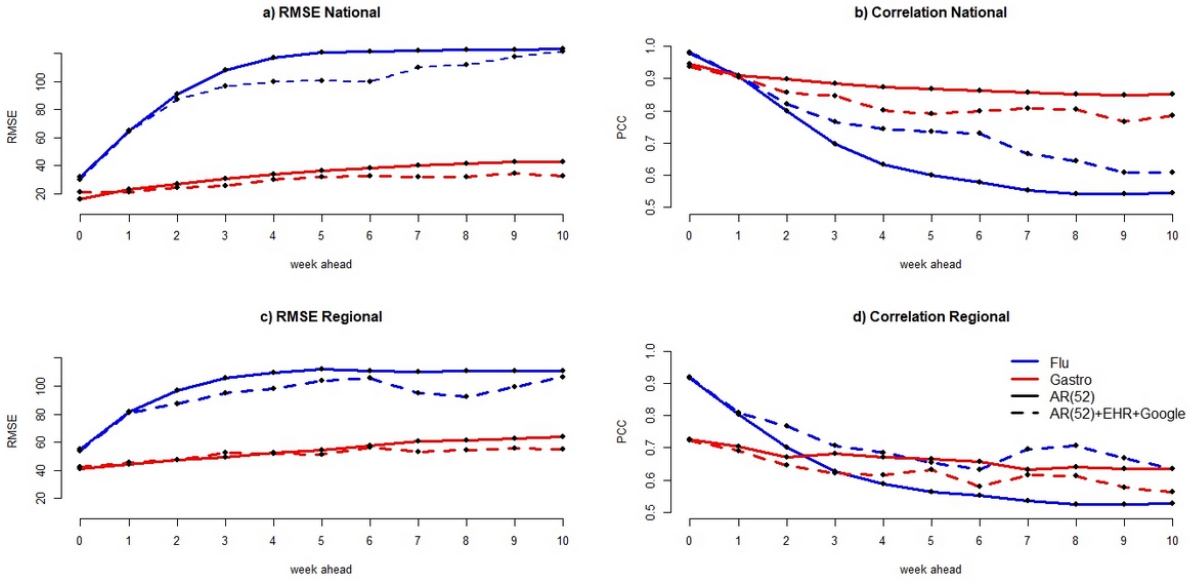


FIGURE 2.17 – Figure 8 de l'article "Gastro bi-sources" [15]. Prévisions jusqu'à 10 semaines avec le modèle Elastic Net pour la grippe et la gastro-entérite. La ligne pleine correspond aux résultats obtenus avec le modèle Elastic Net en utilisant uniquement des données historiques, la ligne pointillée en utilisant toutes les données (historiques, Google et eHOP). Sont en rouge les résultats pour la gastro-entérite et en bleu la grippe. a) Erreurs nationales. b) Corrélations nationales. c) Erreurs régionales. d) Corrélations régionales.

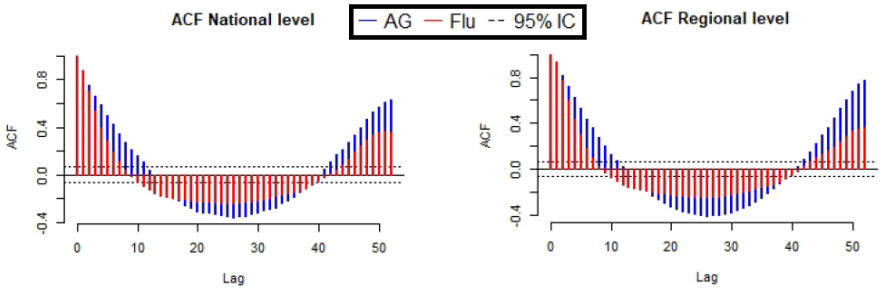


FIGURE 2.18 – Extrait de la Figure 7 de l'article "Gastro bi-sources" [15]. Autocorrelations (ACF) pour la grippe (FLU) et la gastro (AG).

2.5 Perspectives

- L'approche d'apprentissage profond (LSTM) a été testée rapidement pour répondre à un relecteur de notre article [15], et semble prometteuse sur un sujet de recherche à part entière pour adapter le meilleur réseau neuronal à la prévision à long terme. LTSM est lourd en temps de calcul et ne permet pas de détecter facilement les variables les plus importantes pour prédire la gastro-entérite. Néanmoins, certaines études proposent des moyens d'améliorer l'interprétabilité des modèles en perturbant les données d'entraînement ou en utilisant des méthodes basées sur le gradient [70]. La partie la plus lourde du travail sera de développer un autre réseau neuronal plus complexe si nous voulons inclure les sources de données externes telles que les données Google ou Twitter.

- Une autre perspective intéressante est la modélisation conjointe SIR (Susceptible-Infectious-Removed) pour la période épidémique, et apprentissage statistique le reste du temps. Je ne rentrerai pas dans les détails ici des systèmes d'équations différentielles spécifiques que nous devrons développer pour ne pas déflorer le sujet du chapitre suivant, mais me servirai de celle-ci comme transition naturelle.

Ma contribution à la propagation d'épidémie en termes de modélisation de séries temporelles dans un cadre de grande dimension

La thèse de Canelle Poirier, réalisée dans un laboratoire d'informatique médicale où les données réelles sont collectées, où les problématiques médicales émergent, a nécessité des développements de modèles spécifiques de séries temporelles, mais aussi de *machine et deep learning*, avec des comparaisons de performances pour maîtriser les biais d'estimation. Ma contribution, au travers du co-encadrement de cette thèse, a porté sur le développement de modèles mathématiques partant de modèles classiques que nous avons adaptés aux données dont nous disposions et que nous avons collectés. Elle a donné lieu à 4 papiers publiés [11, 12, 13, 15] dans des journaux internationaux d'informatique médicale et dans un journal international scientifique multidisciplinaire (PLOS One), tous avec comité de lecture. Ces articles ont démontré l'intérêt des données massives hospitalières, ainsi que des données Google, Twitter et climatiques dans la modélisation de propagation épidémique, mais aussi les différences de performances des multiples approches possibles de modélisation selon le profil des données. Des perspectives prometteuses en termes de complexité (encore supérieure) de modèles sont envisagées.

Chapitre 3

Propagation d'épidémie : modélisation compartimentale SIR

Ce chapitre repose sur 6 articles dont certains ne sont liés que par l'exploration d'approches couramment utilisées en pharmacocinétique et pharmacodynamie, mais pas en prédiction et caractérisation des épidémies. Je reprendrai quelques éléments d'un article de ma thèse sur les modèles compartimentaux déterministes SIR (Susceptible Infectious Removed), que j'enseigne depuis des années en dernière année d'école d'ingénieurs ENSAI, et citerai aussi la lettre de réponse à auteur sur la cocirculation de virus grippaux qui fut une suite naturelle. Je décrirai ensuite les modèles non linéaires à effet mixtes où je suis co-auteure de trois articles avec Emmanuelle Comets, l'un sur un modèle conjoint à effets mixtes basé sur un modèle compartimental, celui le précédant avec une pré-analyse des données, et un autre co-écrit avec Marc Lavielle de développement du package R saemix (Algorithme d'optimisation des paramètres de modèles non linéaires à effets mixtes par SAEM (Stochastic approximation expectation-maximization). Je parlerai également très rapidement de l'application en croissance tumorale où j'ai optimisé par SAEM un modèle non linéaire à effets mixtes avec interactions. Ces travaux m'ont conduite à tenter une approche non linéaire à effets mixtes basée sur un modèle SIR pour estimer les paramètres moyens et leur variabilité en extrayant d'une série temporelle les épidémies pour les traiter comme des individus statistiques. Je présenterai ce travail préliminaire qui a soulevé plusieurs freins et débouche ainsi sur un ensemble de projets de recherche.

3.1 Principes des modèles compartimentaux

Les modèles compartimentaux sont très largement utilisés dans deux domaines spécifiques : la pharmacocinétique où le devenir du médicament dans l'organisme peut s'envisager de manière dynamique : le corps humain est assimilé à un ensemble de compartiments entre lesquels le médicament peut s'échanger et éventuellement se transformer ; et l'infectiologie où les états de la population face à la maladie à un moment donné définissent à minima les compartiments Susceptibles et Infectieux. Les modèles compartimentaux sont des modèles qui permettent facilement le travail collaboratif entre mathématiciens et médecins de par l'interprétation biologique des paramètres et compartiments. Ce type de modèles de propagation d'épidémie connaît une reprise importante de développement et d'utilisation avec la pandémie de COVID19.

La figure 3.1 montre deux exemples très simples. A droite un modèle à un seul compartiment sché-

matisant le devenir d'une dose de médicament administrée par voie orale, absorbé depuis l'intestin avec une constante d'absorption k_a vers un unique compartiment de volume V, et éliminé avec une constante d'élimination k_e. L'évolution de la quantité Q de médicament dans le sang est alors décrite par une équation différentielle qui dépend de la fonction d'administration du médicament $e(t)$ (orale, intra-veineuse, etc) :

$$\frac{dQ(t)}{dt} = -k_e Q(t) + e(t)$$

A gauche le modèle SIR standard (Susceptible-Infectious-Removed) : les Susceptibles (S) qui peuvent être infectés à tout moment dont une proportion β du nombre maximum de contacts entre les susceptibles et les infectieux devient infectieux chaque jour (β est appelé le taux de contact efficace) ; les Infectieux (I) dont une proportion α se rétablit chaque jour ($\frac{1}{\alpha}$ étant interprété comme la durée de contagiosité), et enfin les Rétablis (ou Retirés du système (R)) qui n'influencent plus la dynamique épidémique.

$$\begin{cases} \frac{dS(t)}{dt} = -\beta S(t)I(t) \\ \frac{dI(t)}{dt} = \beta S(t)I(t) - \alpha I(t) \\ \frac{dR(t)}{dt} = \alpha I(t) \end{cases}$$

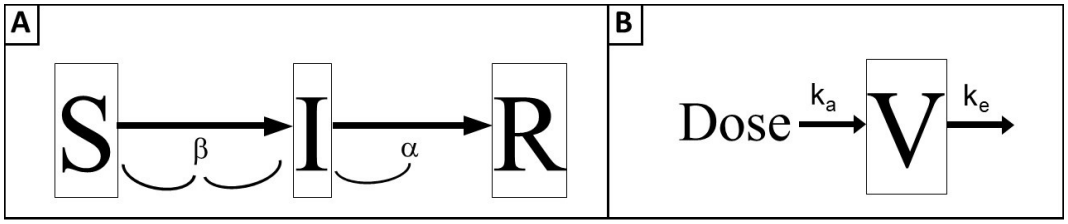


FIGURE 3.1 – Exemples de modèles compartimentaux. A. Modèle de propagation d'épidémie SIR standard. B. Modèle pharmacocinétique à un compartiment avec absorption et élimination du premier ordre.

Ma thèse soutenue en 2004 s'intéressait particulièrement à la co-circulation des virus grippaux au cours d'une même épidémie, avec un premier volet théorique sur la mise en évidence d'une protection croisée lors de co-circulation de plusieurs types antigéniques.

Le modèle construit était basé sur le modèle déterministe SIR standard, modifié de sorte de prendre en compte la co-circulation de 2 souches virales. Pour explorer les effets temporels de la diffusion d'une première souche (Virus 1) émergeant au temps t_1 et celle d'une seconde souche (Virus 2) émergeant au temps $t_2 = t_1 + \tau$, une série de simulations a été réalisée pour différents décalages τ et différents niveaux de protection croisée ($\eta_{1/2}$ et $\eta_{2/1}$).

Les individus du compartiment S_1S_2 sont susceptibles aux 2 souches, les I_1S_2 infectieux infectés par le virus 1 mais encore susceptibles au virus 2 (réciproquement pour les S_1I_2), les R_1S_2 ne sont plus infectieux "virus 1" mais encore susceptibles au virus 2 (réciproquement pour les S_1R_2), les R_1I_2 ne sont plus infectieux "virus 1" et sont maintenant infectieux infectés par le virus 2 (réciproquement pour les I_1R_2), les I_1I_2 sont infectieux infectés par les 2 souches, et les R_1R_2 ne participent plus au processus de contamination. Les paramètres β_1 et β_2 représentent les taux de contact efficaces pour l'infection respectivement au virus 1 et 2. Les paramètres α_1 et α_2 correspondent à la vitesse moyenne de guérison (ou plus exactement de sortie des compartiments Infectieux).

Le système d'équations différentielles ci-après est notre modèle structural.

$$
\begin{cases}
\frac{dS_1S_2(t)}{dt} = -S_1S_2(t)(\beta_1(I_1S_2(t)+I_1R_2(t))+\beta_2(S_1I_2(t)+R_1I_2(t))+(\beta_1+\beta_2)I_1I_2(t)) \\
\frac{dR_1S_2(t)}{dt} = \alpha_1 I_1S_2(t) - R_1S_2(t)\beta_2(1-\eta_{2/1})(S_1I_2(t)+R_1I_2(t)+I_1I_2(t)) \\
\frac{dS_1R_2(t)}{dt} = \alpha_2 S_1I_2(t) - S_1R_2(t)\beta_1(1-\eta_{1/2})(I_1S_2(t)+I_1R_2(t)+I_1I_2(t)) \\
\frac{dI_1S_2(t)}{dt} = S_1S_2(t)\beta_1(I_1S_2(t)+I_1R_2(t)+I_1I_2(t)) - \alpha_1 I_1S_2(t) \\
\qquad\qquad - I_1S_2(t)\beta_2(S_1I_2(t)+R_1I_2(t)+I_1I_2(t)) \\
\frac{dS_1I_2(t)}{dt} = S_1S_2(t)\beta_2(S_1I_2(t)+R_1I_2(t)+I_1I_2(t)) - \alpha_2 S_1I_2(t) \\
\qquad\qquad - S_1I_2(t)\beta_1(I_1S_2(t)+I_1R_2(t)+I_1I_2(t)) \\
\frac{dR_1I_2(t)}{dt} = R_1S_2(t)\beta_2(1-\eta_{2/1})(S_1I_2(t)+R_1I_2(t)+I_1I_2(t)) - \alpha_2 R_1I_2(t) \\
\frac{dI_1R_2(t)}{dt} = S_1R_2(t)\beta_1(1-\eta_{1/2})(I_1S_2(t)+I_1R_2(t)+I_1I_2(t)) - \alpha_1 I_1R_2(t) \\
\frac{dI_1I_2(t)}{dt} = I_1S_2(t)\beta_2(S_1I_2(t)+R_1I_2(t)+I_1I_2(t)) + S_1I_2(t)\beta_1(I_1S_2(t)+I_1R_2(t) \\
\qquad\qquad + I_1I_2(t)) - \frac{\beta_1\alpha_1+\beta_2\alpha_2}{\beta_1+\beta_2} I_1I_2(t) \\
\frac{dR_1R_2(t)}{dt} = \alpha_1 I_1R_1(t) + \alpha_2 R_1I_2(t) + \frac{\beta_1\alpha_1+\beta_2\alpha_2}{\beta_1+\beta_2} I_1I_2(t)
\end{cases}
$$

Par simulations et analyse de sensibilité, nous avons montré qu'une protection croisée de 50% est suffisante pour expliquer pourquoi nous observons un seul pic annuel de syndrome grippal dans les pays tempérés [1].

Une réponse à auteur a également été publiée sur un article discutant de l'immunité croisée entre sous-types de virus grippaux [2].

3.2 Modèles non linéaires à effets mixtes

En recherche clinique, nous avons également souvent des "séries temporelles" mais à l'échelle individuelle. On parle alors de données longitudinales ou de mesures répétées. Ces données présentent souvent une structure hiérarchique, avec des corrélations introduites par les mesures répétées sur le même individu. Ces corrélations peuvent être gérées en modélisant l'évolution d'un processus avec le temps et en supposant des paramètres spécifiques à l'individu, pour tenir compte des différences inter-individuelles. Les modèles utilisés pour décrire la dynamique des processus biologiques sont souvent non linéaires en ce qui concerne les paramètres impliqués, et les outils statistiques appropriés dans ce contexte sont les modèles non linéaires à effets mixtes [71]. Les données longitudinales jouent par exemple un rôle important dans le processus de développement des nouveaux médicaments, et l'analyse pharmacocinétique fait partie intégrante du dossier d'enregistrement soumis à l'autorité sanitaire pour l'approbation de nouveaux médicaments [72]. Il est également de plus en plus utilisé pour adapter le traitement médicamenteux et guider l'adaptation de la dose dans des populations particulières, comme par exemple dans le traitement par voie rénale de patients ou enfants ayant des facultés affaiblies.

3.2.1 Package R saemix

Soit Y la variable à modéliser, qui peut être par exemple en pharmacocinétique, une concentration. Le modèle s'écrit :

$$y_{ij} = f(x_{ij},\psi_i) + g(x_{ij},\psi_i,\xi)\epsilon_{ij}, \quad 1 \le i \le N, \quad 1 \le j \le n_i$$

où $y_{ij} \in \mathbb{R}$ est la j$^{\text{ème}}$ valeur observée du sujet i,
• N le nombre de sujets, n_i le nombre d'observations du sujet i,
• f le modèle structural,
• x_{ij} les valeurs observées pour les régresseurs ou variables de design (pour les modèles pharmacocinétiques (PK) ou pharmacocinétiques-pharmacodynamiques (PK-PD), x est souvent la dose ou le temps),

• ψ_i les paramètres individuels, inconnus. $\psi_i = h(C_i\mu + \eta_i)$,
avec $\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Omega)$, les vecteurs aléatoires, inconnus, correspondant aux effets aléatoires et
C la matrice des covariables,
h la fonction de transformation du vecteur gaussien,
μ le vecteur des paramètres de population, inconnu, correspondant aux effets fixes,
• $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ les erreurs,
• g la fonction définissant le modèle d'erreurs avec les paramètres ξ.

Ce modèle implique donc un certain nombre d'hypothèses sur :
— le modèle structural f qui dépendra des variables de régression comme le temps ou la dose, mais aussi de ψ_i les paramètres individuels inconnus, tenant compte des paramètres fixes de population μ, des effets des covariables C_i, et d'effets aléatoires η_i gaussiens dont le paramètre Ω sera à estimer.
— d'autres hypothèses notamment sur le modèle d'erreur sont à poser au travers de la fonction g.

Dans ce modèle, on se retrouve finalement dans le cas de données incomplètes, puisque les données complètes du modèle sont les y_{ij} observées et les ψ_i non observées propres à l'individu. Il s'agit donc d'estimer le jeu de paramètres inconnus θ comprenant μ, Ω, et ξ, avec ξ correspondant aux paramètres du modèle d'erreur.

L'estimation des paramètres des modèles peut être effectuée en utilisant des approches de maximum de vraisemblance et les algorithmes d'optimisation peuvent être d'une importance capitale. En raison de l'importance de ces modèles en pharmacocinétique, les premières méthodes d'estimation ont été développées en pharmacométrie. Dans ce contexte, un logiciel dédié, appelé NONMEM, a été développé dans les années 70, celui-ci traite les caractéristiques spécifiques des données pharmacologiques telles que le schéma posologique et d'autres variables mesurées pendant le traitement [73]. Le premier des algorithmes implémentés dans ce logiciel s'appuyaient sur la linéarisation du modèle pour obtenir une approximation de la vraisemblance, qui ne peut pas être calculée facilement dans les modèles non linéaires à effets mixtes. Cette approximation est ensuite maximisée par minimisation itérative de Newton-Raphson, un algorithme à usage général impliquant le gradient de la fonction à optimiser. Différentes approximations de la vraisemblance ont été proposées, avec par exemple des méthodes où la linéarisation se fait au niveau des estimations individuelles à chaque itération [74]. Ces méthodes d'estimation ont également été mises en œuvre dans les logiciels statistiques grand public tels que SAS (PROC NLMIXED ; SAS Institute Inc. 2013) et R (R Core Team 2017), où le package nlme [75] fait maintenant partie de l'installation de base. Les méthodes basées sur la linéarisation présentent cependant des lacunes à la fois statistiques et pratiques. Premièrement et surtout, elles ne convergent pas vers les estimations du maximum de vraisemblance. Le biais est généralement mineur pour les effets fixes, mais les composantes de variance peuvent être significativement biaisées, en particulier avec une grande variabilité interindividuelle. Il a été démontré que cela augmentait l'erreur de type I des tests de vraisemblance [76, 77], avec le risque de construire de mauvais modèles. Deuxièmement, ces méthodes souffrent de biais sévères lorsqu'elles sont appliquées à des données non continues, comme l'ont montré Molenberghs et Verbeke en 2005 [78], alors qu'il a été démontré que les algorithmes stochastiques fournissent des estimations non biaisées [79]. Dans la pratique, les algorithmes basés sur la linéarisation présentent, en plus, des problèmes de convergence et peuvent être difficiles à utiliser avec des modèles complexes [80]. Pour palier à cela, de nouveaux et puissants algorithmes d'estimation ont donc été proposé afin d'estimer les paramètres des modèles non linéaires à effets mixtes [81].
L'alternative à la linéarisation du modèle est de calculer la vraisemblance par approximation numérique ou statistique qui préservent les propriétés statistiques des estimateurs du maximum de vraisemblance.

Un exemple est l'approximation de Laplace, qui est équivalente à la Quadrature de Gauss avec un point pris sur le domaine d'intégration, et a été implémentée dans NONMEM. Dans R, le package lme4 utilise cette approximation [82] avec un moindre carré pénalisé comme algorithme d'estimation. Une alternative puissante à la minimisation basée sur les algorithmes du gradient est l'algorithme EM (algorithme itératif développé dans le contexte des données manquantes [83]). L'algorithme SAEM pour Stochastic Approximation Expectation Maximization, combinant une approximation stochastique de la vraisemblance avec un algorithme EM, s'est avéré très efficace pour les modèles non linéaires à effets mixtes, convergeant rapidement vers les estimateurs du maximum de vraisemblance [84].

L'écriture de la vraisemblance de notre modèle non linéaire à effets mixtes est donc :

$$l(\theta;y) = \prod_{i=1}^{N} l(\theta;y_i) = \prod_{i=1}^{N} p(y_i|\theta) = \prod_{i=1}^{N} \int_D p(y_i|\eta_i,\theta)p(\eta_i|\theta)d\eta_i)$$

où D est la distribution des paramètres individuels.

Cette vraisemblance n'a pas d'expression analytique lorsque f est non linéaire. Pour contourner ce problème, deux approches principales peuvent être utilisées. La première approche implique une linéarisation du modèle [74] ou de la vraisemblance [85], tandis que la deuxième approche utilise le l'approximation numérique [75] ou stochastique [86] pour calculer la vraisemblance. La vraisemblance approchée est donc ensuite maximisée grâce aux algorithmes tels que le quasi-Newton, ou par des algorithmes EM [83], où les paramètres individuels inconnus sont traités comme des données manquantes. En modèles non linéaires à effets mixtes, l'étape E à l'itération k de l'algorithme EM consiste à calculer l'espérance conditionnelle du log de la probabilité d'observer les données complètes :

$$Q_k(\theta) = E(\log(p(y,\psi;\theta)|y,\theta_{k-1})$$

et l'étape M consiste à calculer la valeur k qui maximise $Q_k(\theta)$.

Quand la fonction f ne dépend pas linéairement des effets aléatoires, l'étape E ne peut pas être réalisée normalement. La version d'approximation stochastique de l'algorithme EM standard [84], consiste à remplacer l'habituelle étape E de l'EM par une procédure stochastique. À l'itération k de SAEM, l'algorithme procède comme suit :

— Etape Simulation : dessine $\psi^{(k)}$ à partir de la distribution conditionnelle $p(.|y;\theta_k)$.

— Approximation stochastique : mise à jour de $Q_k(\theta)$ selon

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\log(p(y,\psi^{(k)};\theta) - Q_{k-1}(\theta)).$$

— Etape Maximisation : mise à jour de θ_k selon

$$\theta_{k+1} = \text{argmax}_\theta \, Q_k(\theta).$$

Cet algorithme a été implémenté dans le logiciel Monolix [81] qui a connu une croissance forte à partir de là, il est également disponible dans la boîte à outils Statistiques de MATLAB (nlmefitsa.m ; [87]), mais aussi dans NONMEM version 7.
Nous avons implémenté l'algorithme SAEM dans le logiciel R via le package saemix [88]. Le package utilise le système de classe S4 de R pour fournir un système d'entrée et de sortie convivial. Il fournit des résumés des résultats, les estimations de paramètres individuels, les erreurs-types (obtenues à l'aide d'un calcul de la matrice d'information de Fisher), des Tests de Wald pour les effets fixes, et de nombreux graphiques diagnostiques, comprenant les Visual Predictive Check (VPC) et les Normalized Prediction Distribution Errors (npde) ([89]. Les graphiques de diagnostic peuvent être adaptés aux préférences de

l'utilisateur en définissant un certain nombre d'options, et sont facilement exportés vers un fichier. Dans l'article [9], nous présentons d'abord les modèles statistiques, puis nous décrivons les caractéristiques du package en l'appliquant à plusieurs exemples. Nous concluons par une étude de simulations évaluant les performances de saemix et ses caractéristiques de fonctionnement.

J'ai présenté notre package saemix aux Rencontres R à Bordeaux en 2012 [90], après qu'il ait été présenté au congrès PAGE en 2011 [91].

3.2.2 Applications en pharmacodynamie

Étude sur la croissance tumorale

J'ai réalisé l'analyse des données d'une étude de l'efficacité du sunitinib (TKI, l'Inhibiteur de Tyrosine Kinase) en association avec les antagonistes des récepteurs de type 1 à l'angiotensine-II (ARA2, le telmisartan) sur le modèle murin de xénogreffe de carcinome rénal à cellules claires (article [5]). Des souris tests ont été injectées ensuite avec 10 millions de cellules tumorales en sous-cutané dans le flanc. La mise au point du modèle a permis de développer des tumeurs en 5 semaines chez des souris avec une moyenne tumorale de 200 mm^3. L'étape suivante a consisté à injecter 40 souris, de les diviser en 4 groupes, puis de tester notre hypothèse de potentialisation du sunitinib par le telmisartan sur un cycle de traitement de 4 semaines. Le plan expérimental est un plan factoriel 2×2 conduisant à 4 groupes selon le tableau 3.1 avec 10 souris dans chaque groupe.

	Pas de TKI	TKI
Pas de ARA2	**Aucun traitement**	**TKI seul**
ARA2	**ARA2 seul**	**TKI et ARA2**

TABLE 3.1 – Plan expérimental : plan fatoriel 2×2. Le groupe "Aucun traitement" a reçu un placebo (le diluant classiquement utilisé pour les médicaments : le diméthylsulfoxyde DMSO)

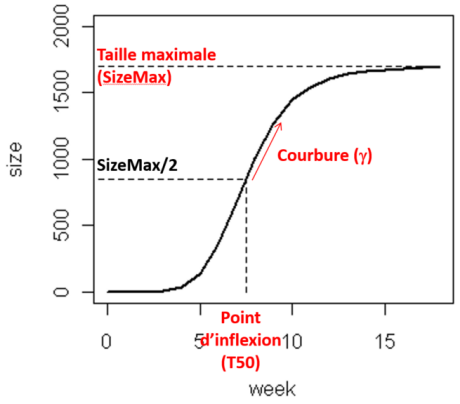


FIGURE 3.2 – Illustration des 3 paramètres du modèle sigmoïde : Taille maximale, courbure, et abscisse du point d'inflexion. Ici $SizeMax = 1700, T50 = 7.5, \gamma = 6$.

A l'issue de cette période, les animaux étaient sacrifiés. Le sang et la tumeur étaient prélevés pour analyse biologique (ELISA), histologique (IHC), et protéique (Western Blot). La méthodologie et les résultats sont présentés dans l'article [5]. Il en ressort que l'association des deux traitements engendrait plus de nécrose tumorale (p= 0,038), une réduction de la densité microvasculaire centrale (p= 0,038), et une réduction du VEGF circulant (la protéine "facteur de croissance de l'endothélium vasculaire"). En revanche, aucune différence n'était notée concernant la prolifération tumorale et l'apoptose tumorale en analyse IHC comme en Western blot.

J'ai construit un modèle non linéaire à effets mixtes pour étudier l'effet dans le temps des deux traitements et de leur interaction sur la croissance tumorale, et utilisé notre package saemix.
Le modèle structural f est un modèle sigmoïde à 3 paramètres $\theta = (Sizemax, T50, \gamma)$ dont l'interprétation se comprend facilement avec le schéma 3.2.

Soit Y la croissance tumorale. Le modèle s'écrit : $y_{ij} = f(t_{ij}, \psi_i) + g(x_{ij}, \psi_i, \xi)\epsilon_{ij}$ avec le modèle structural sigmoïde f dépendant de $\theta = (Sizemax, T50, \gamma)$:

$$f(\theta_i, t_{ij}) = \frac{Sizemax_i (t_{ij})^{\gamma_i}}{(T50_i)^{\gamma_i} + (t_{ij})^{\gamma_i}}$$

et ψ_i les paramètres individuels, inconnus $\psi_i = h(C_i\mu + \eta_i)$
avec

- $\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Omega)$, les vecteurs aléatoires, inconnus, correspondant aux effets aléatoires,
- C la matrice des covariables avec les deux traitements,
- h la fonction de transformation du vecteur gaussien (ici choix de la fonction log)
- μ le vecteur des paramètres de population, inconnu, correspondant aux effets fixes, de sorte que

$$\begin{cases} \theta_i = \mu \exp(\eta_i) & \text{si Aucun traitement} \\ \theta_i = \mu \exp(\beta_1 \eta_i) & \text{si ARA2 seul} \\ \theta_i = \mu \exp(\beta_2 \eta_i) & \text{si TKI seul} \\ \theta_i = \mu \exp((\beta_1 + \beta_2 + \beta_{12})\eta_i) & \text{si TKI et ARA2} \end{cases}$$

et $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ les erreurs,
g la fonction définissant le modèle d'erreurs avec les paramètres ξ (ici choix de $g = a + bf$ et $\xi = (a, b)$).

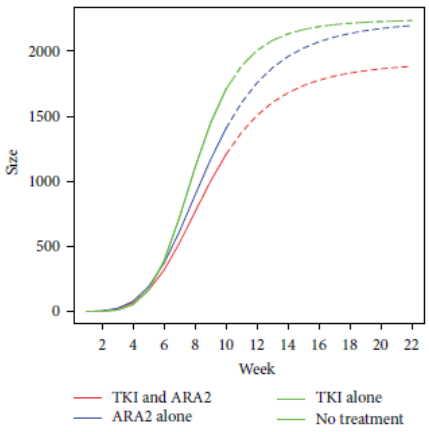


FIGURE 3.3 – Figure 2 (b) de l'article [5]. Courbes cinétiques avec prolongation jusqu'à 22 semaines des 4 groupes de traitement.

Nous avons montré une vitesse de croissance ralentie avec ARA2, et le meilleur modèle après sélection de covariables sur les différents paramètres laisse entrevoir une tendance, cependant non significative (p=0.069), à une diminution de la taille « finale » quand on ajoute TKI à ARA2 (figures 3.3 et 3.4).

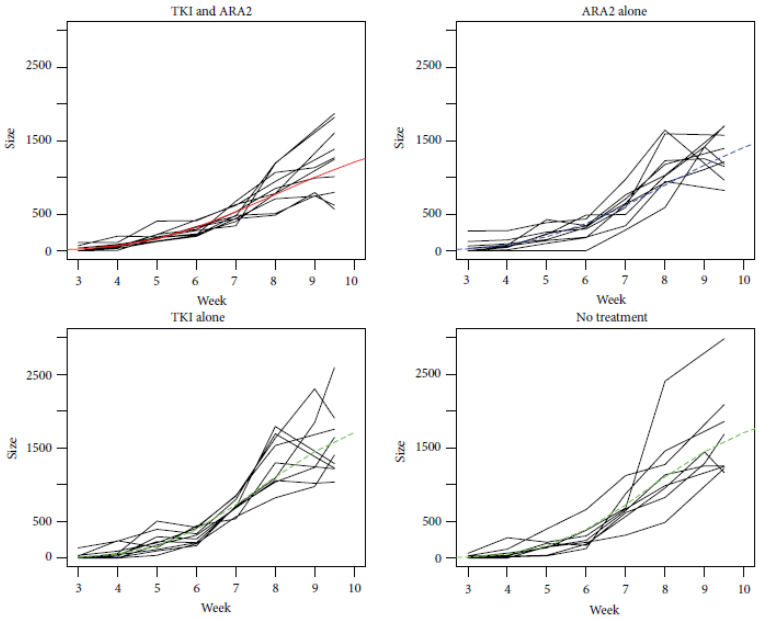


FIGURE 3.4 – Figure 2 (a) de l'article [5]. Cinétiques de la croissance tumorale des 4 groupes de traitement.

Étude HEPMEN

En rentrant moins dans les détails que sur l'étude précédente, mais pour montrer un cas de modèle non linéaire à effets mixtes avec un modèle structural compartimental géré par des équations différentielles, je présente ici le modèle développé au cours du stage, que j'ai co-encadré, de master 2 Ingénierie mathématiques option statistique de l'Université de Nantes, et qui a été publié dans l'article [6] à la suite d'un premier article plus descriptif de cette recherche clinique originale [8]. Au vue de la littérature, on a pu supposer que le fer régulait la synthèse d'hepcidine et que l'hepcidine régulait l'absorption du fer [92]. Ainsi, une augmentation de fer va induire une augmentation de synthèse d'hepcidine et inversement. Nous avons modélisé le fer et l'hepcidine conjointement à l'aide d'un modèle de "turnover" (réponse indirecte) reflétant les faits que l'hepcidine n'agit pas directement sur le fer mais sur son absorption et que le fer n'agit pas sur l'hepcidine mais sur sa synthèse (schéma 3.5).

Nous avons utilisé les données recueillies dans l'étude HEPMEN pendant le cycle menstruel de 90 femmes en bonne santé (6 prises de sang tout au long du cycle). Un schéma général a été observé pour l'hepcidine et le fer, avec une diminution initiale pendant la menstruation, suivie d'un rebond et d'une stabilisation au cours de la seconde moitié du cycle.

Le modèle conjoint développé est le suivant :

$$\begin{cases} \frac{d\mathrm{Ir}(t)}{dt} = \mathrm{ksyn}_I(t) - \mathrm{kout}_I.\mathrm{Ir}(t) \\ \frac{d\mathrm{He}(t)}{dt} = \mathrm{ksyn}_H(t)(1 + \alpha(\mathrm{Ir}(t) - \mathrm{Ir}_0)) - \mathrm{kout}_H.\mathrm{He}(t) \end{cases}$$

démarrant aux conditions initiales :

$$\begin{cases} \mathrm{Ir}(t=0) = \frac{\mathrm{ksyn}_{I_0}}{\mathrm{kout}_{I_0}} \\ \mathrm{He}(t=0) = \frac{\mathrm{ksyn}_{H_0}}{\mathrm{kout}_{H_0}} \end{cases}$$

où $t = 0$ est le début du cycle ; $\mathrm{ksyn}_I, \mathrm{kout}_I, \mathrm{ksyn}_H$et kout_H, varient avec t selon la figure 3.5 (à droite) et *dloss* désigne la durée des règles de chaque femme et a été fixé aux valeurs individuelles observées, tandis que *drel* est une constante estimée, supposée positive et sans variabilité interindividuelle. Avec l'hypothèse qu'une autre perturbation intervient au cours du cycle, on suppose que, suite à la perte de fer causée par les règles, le corps va relarguer une quantité supplémentaire de fer afin de compenser. On représente ce phénomène à l'aide de la constante krel présente sur la période [2 ;2+drel]. Enfin, dcycle a été fixé à la longueur du cycle pour chaque femme.

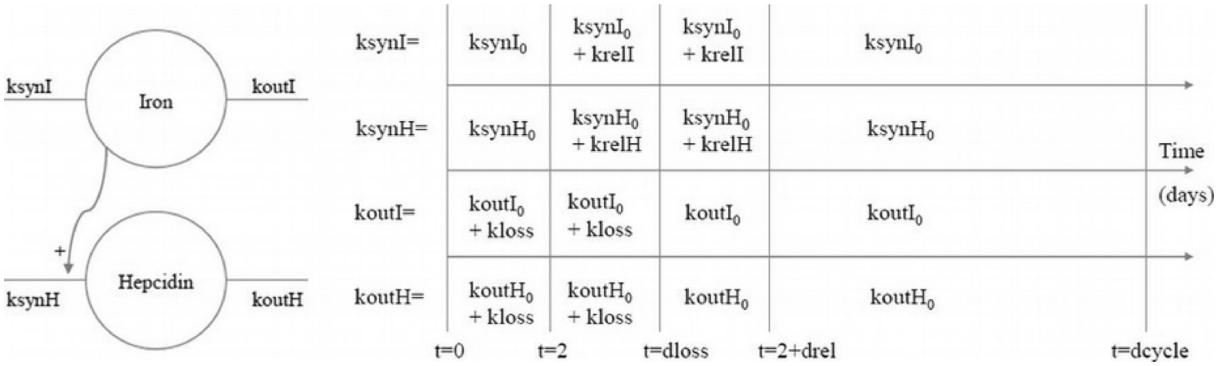


FIGURE 3.5 – Figure 3 de l'article [6]. Modèle pour le fer et l'hepcidine chez les femmes non ménopausées (à gauche), et valeurs des paramètres contrôlant la réponse indirecte des deux molécules selon le moment du cycle menstruel (à droite)

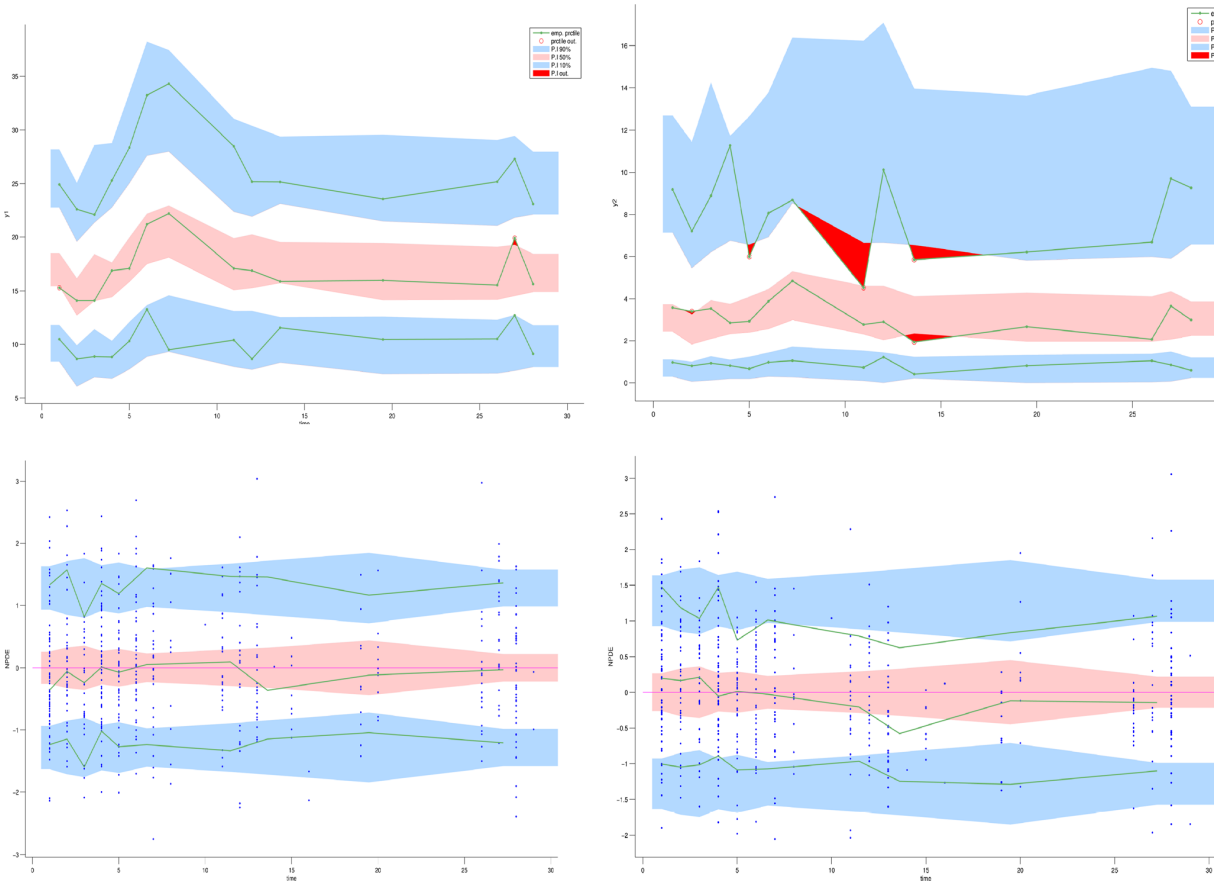


FIGURE 3.6 – Figure 5 de l'article [6]. Diagnostic VPC (en haut) et graphiques des npde (en bas), tels qu'obtenus avec le logiciel Monolix pour le fer (à gauche) et l'hepcidine (à droite). Les lignes pleines représentent les 10e, 50e et 90e percentiles empiriques des concentrations mesurées pour les VPC, et ceux des résidus pour les graphiques npde. Les zones colorées représentent l'intervalle de prédiction à 90% associé aux 10e, 50e et 90e percentile théorique dans le premier cas et l'intervalle de confiance dans le second.

Grâce à ce modèle bien adapté et correctement ajusté, au vu des graphiques diagnostiques de la figure 3.6, nous avons pu montrer que le fer a stimulé la libération d'hepcidine. Plusieurs covariables, y compris la contraception, la quantité de sang perdue et la ferritine, ont influé sur les paramètres. Ce modèle conjoint a apporté une compréhension fondamentale du phénomène.

3.2.3 Perspectives d'application des modèles non linéaires à effets mixtes aux épidémies

Application aux données grippales

Mon expérience aujourd'hui dans les modèles non linéaires à effets mixtes au sein de l'équipe PK-PD du CIC 1414 m'a décidée à coupler différentes approches dans un but de description clinique des épidémies et de quantification de la variabilité des paramètres d'un modèle structural Susceptible-Infectious-Removed sophistiqué tenant également compte de régresseurs et de covariables explicatives spécifiques à chaque épidémie (souches virales, vacances scolaires, météo,...).

La dérive antigénique, les clusters de virus circulant durant un à cinq ans, les variabilités de la couverture et efficacité vaccinales, et la co-circulation de types ou sous-types de virus font que les épidémies se suivent mais ne se ressemblent pas. Certains modélisateurs ont pu par exemple appliquer un saut non constant sur le nombre de susceptibles en début d'épidémie même en modélisation Susceptible-Exposed-Infectious-Removed-Susceptible (SEIRS) sur plusieurs années [93]. Finalement en estimant le pourcentage de susceptibles au début de chaque épidémie (p_{S_0}), la modélisation de chacune d'entre elles peut se faire indépendamment des autres. Ainsi est né mon intérêt pour l'étude des paramètres épidémiologiques et leur variabilité inter-épidémique par un modèle non linéaire à effet mixte, en pleine conscience des limites statistiques de considérer l'indépendance des individus statistiques (l'épidémie) alors que p_{S_0} repose sur une équation complexe mais dépendante du passé.

La construction du modèle structurel est fortement basée sur le modèle de Truscott-Ferguson [94], mais aussi avec des compléments de la littérature [93, 95, 96, 97, 98, 99, 100].

Soit Y l'incidence grippale. Le modèle s'écrit :

$$y_{ij} = f(t_{ij}, \psi_i) + g(x_{ij}, \psi_i, \xi)\epsilon_{ij}$$

avec le modèle structural suivant :

$$f(\theta_i, t_{ij}) = \frac{pconsult.E}{pctEduree.\theta_I}$$

où E et I sont gérés par le système d'équations différentielles suivant :

$$\left\{ \begin{array}{l} \frac{dS}{dt} = -\lambda S \\ \frac{dE}{dt} = \lambda S - \frac{E}{pctEduree.\theta_I} \\ \frac{dI}{dt} = \frac{E}{pctEduree.\theta_I} - \frac{I}{\theta_I} \end{array} \right.$$

avec dépendant du temps la force d'infection λ :

$$\left\{ \begin{array}{l} R0 = \frac{(1+R0_{base})(1-\alpha_{vac}.X^{(Vac)})\exp(\beta_1.\frac{X^{(Temp)}-\bar{X}^{(Temp)}}{\sigma_{X^{(Temp)}}})}{pS0} \\ \lambda = \frac{R0}{\theta_I}\frac{I^\nu}{N} \end{array} \right.$$

et l'initialisation des 3 compartiments à :

$$\left\{ \begin{array}{l} S(t=0) = pS0.N \\ E(t=0) = \frac{I_0}{pctEduree} \\ I(t=0) = I_0 \end{array} \right.$$

avec $N = 100000$ et les paramètres à estimer :
❍ $\theta = (pS0, R0_{base}, \theta_I, I_0, \nu, pctEduree, pconsult, \alpha_{vac}, \beta_1, \beta_{cov})$
❍ ψ_i les paramètres individuels, inconnus. $\psi_i = h(C_i\mu + \eta_i)$, avec
✦ $\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Omega)$, les vecteurs aléatoires, inconnus, correspondant aux effets aléatoires et
✦ C la matrice des covariables qui sont :
- l'incidence observée à t=0,
- la présence d'un virus de type B,
- la cocirculation de types ou sous-types et
- le moment où l'épidémie a démarré (en semaine d'année scolaire et non civile puisque dans l'hémisphère Nord les épidémies démarrent entre septembre et avril).

✦ h la fonction choisie log ou logit de transformation du vecteur gaussien selon si l'on souhaite par exemple contraindre les paramètres entre 0 et 1 (pour les pourcentages),

✦ μ le vecteur des paramètres de population, inconnu, correspondant aux effets fixes (impactés par β_{cov} selon les covariables),

$\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ les erreurs,

g la fonction définissant le modèle d'erreurs avec les paramètres ξ. Ici, après essai de différentes fonctions, j'ai choisi $g = bf$ et $\xi = b$ (modèle d'erreurs proportionnel).

Afin de modéliser certains aspects de l'hétérogénéité de la population participant à la propagation de la grippe, certains auteurs utilisent un terme puissance ν sur la population de Susceptibles S(t) [101]. De nombreux modèles ont été comparés étapes par étapes pour minimiser le critère BIC (Bayesian information Criterion) pour choisir les meilleurs modèles parmi SEIR, S$^{\nu}$EIR, SEI$^{\nu}$R avec 3 combinaisons de variabilités des paramètres *pctEduree* et *R0base* (sur 5 combinaisons *pconsult* et θ_I fixés ou pas), ont également été testés le rajout de 2 classes d'âge qui n'a pas été retenu, et des régresseurs sur le nombre de reproduction :
- 3 régresseurs météo (température à Nantes, Pluviométrie, ensoleillement), nous avons retenu uniquement la température, comme indiqué dans l'écriture du modèle.
- 3 régresseurs vacances : « oui/non », « toussaint/noel/hiver/printemps » ou « 1$^{\text{ères}}$/2$^{\text{èmes}}$ », nous avons retenu la variable binaire « oui/non », comme indiqué là aussi dans l'écriture du modèle.
A partir des 6 meilleurs modèles, une sélection des covariables pas à pas descendante manuelle a été réalisée en minimisant le critère BIC (effet du virus type B, de la cocirculation de types ou sous-types grippaux,... sur chaque paramètre). Notons qu'il y a des contraintes sur l'optimisation des paramètres pour un ordre de grandeur raisonnable.

	parameter	s.e. (s.a.)	r.s.e.(%)	p-value
$pS0_{pop}$	0.406	0.072	18	
$\beta_{pS0_{semdebutepi}}$	-0.0577	0.017	30	0.00076
$R0base_{pop}$	0.8	0.11	13	
$\beta_{R0base_{virusB_{oui}}}$	-0.248	0.046	18	5.9e-008
$\theta_{I_{pop}}$	0.34	-	-	
$I0_{pop}$	15.1	1.4	9	
$\beta_{I0_{virusB_{oui}}}$	0.614	0.13	21	3.3e-006
ν_{pop}	0.958	0.01	1	
$pctEduree_{pop}$	0.596	0.0082	1	
$\beta_{pctEduree_{cocircu_{oui}}}$	0.318	0.093	29	0.00063
$pconsult_{pop}$	0.5	-	-	
$\alpha_{vac_{pop}}$	0.116	0.018	16	
$\beta_{1_{pop}}$	-0.0338	0.002	6	
Ω_{pS0}	0.404	0.063	15	
Ω_{R0base}	0.0962	0.021	22	
Ω_{I_0}	0.272	0.055	20	
$\Omega_{\alpha_{vac}}$	0.68	0.14	20	
b	0.198	0.011	5	

TABLE 3.2 – Estimation des paramètres de population avec le modèle final.

Le modèle final (table 3.2) estime un pourcentage initial moyen de susceptibles ($pS0_{pop}$) à 40% qui diminue quand le début d'épidémie tarde ($\beta_{pS0_{semdebutepi}}$ négatif), laissant éventuellement imaginer la possibilité de vaccinations tardives. De plus ce modèle comprend une variabilité inter-épidémique de ce paramètre (en estimant Ω_{pS0}) qui permet donc de "décorréler" les épidémies en tenant compte d'un effet

propre qu'il soit ou non dû au passé.
Le paramètre $R0base_{pop}$ est estimé à 0.8 et conduit donc à un paramètre R0 pouvant être interprété un peu comme un nombre de reproduction de base de 1.8 dans une population 100% susceptible hors période de vacances scolaires et pour une température moyenne. Une variabilité inter-épidémique de ce paramètre a été appliquée permettant d'envisager une contagiosité différente chaque année. Quand l'épidémie est due à une souche grippale de type B, ce paramètre est diminué ($\beta_{R0base_{virusB_{oui}}}$ négatif), ce résultat est cohérent avec la connaissance que les virus de type B sont en général moins contagieux.
Le terme $\theta_{I_{pop}}$ est couramment interprété comme la durée de contagiosité [102, 103], mais aussi comme la moyenne de l'intervalle de génération [104, 105]. Il a été estimé à 2.4 jours sur les épidémies françaises de 1984 à 2007 dans l'article [106]. Nous l'avons fixé à cette valeur. Nous avons fixé un deuxième paramètre, le pourcentage de consultation $pconsult_{pop}$ à 50% d'après ce même article.
Les paramètres $I0_{pop}, \beta_{I0_{virusB_{oui}}}$ et Ω_{I_0} sont dans notre modèle parce que nous avons tronqué les épidémies uniquement aux semaines dites épidémiques, l'interprétation des estimations n'a pas d'intérêt épidémiologique.
Le paramètre ν_{pop} est estimé à 0.958 en cohérence avec l'article de Hickmann [101].
Le paramètre $pctEduree_{pop}$ est estimé à 0.59, s'interprétant comme un peu moins d'un jour et demi passés dans le compartiment E infecté mais pas encore contagieux (puisqu'il est à multiplier par $\theta_{I_{pop}}$), et $\beta_{pctEduree_{cocircu_{oui}}}$ étant positif il semblerait que la cocirculation de virus grippaux augmenterait un peu cette période de latence.
Le paramètre $\alpha_{vac_{pop}}$ estimé à 0.116 fait diminuer le nombre de reproduction de base pendant les périodes de vacances scolaires différemment selon les années puisqu'une variabilité inter-épidémique a été appliquée sur ce paramètre ($\Omega_{\alpha_{vac}}$).
Enfin $\beta_{1_{pop}}$ est négatif et fait augmenter le nombre de reproduction de base quand la température diminue, pouvant s'expliquer par des rassemblements plus importants en lieux confinés quand il fait froid et un virus qui n'aime pas la chaleur.

Le graphique des Visual Predictive Check (VPC) constitue un outil diagnostique qui a fait ses preuves, il permet la comparaison entre les données observées et des intervalles de prédiction. Le principe est la

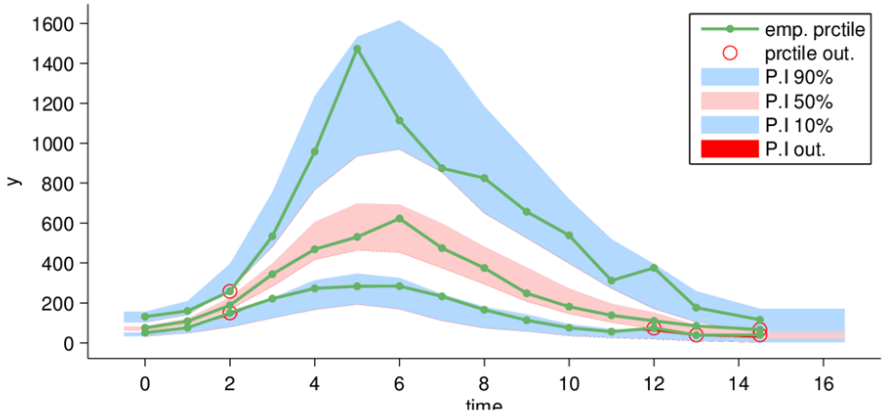


Figure 3.7 – Les VPC des 24 épidémies.

simulation de jeux d'observations sous le modèle qui permet la construction de l'enveloppe où devrait se trouver un pourcentage donné d'observations. Sur la figure 3.7, les données observées sont représentées par les points bleus. Les traits continus sont les $2.5^{\text{ème}}$, $50^{\text{ème}}$ et $97.5^{\text{ème}}$ percentile des données observées.

Par défaut, 1000 jeux sont simulés. La courbe pointillée rouge correspond à la médiane des 1000 médianes, entourée de l'intervalle à 95% de prédiction de cette médiane. De même en bleu pour les percentiles 2.5 et 97.5. Les traits continus se trouvent dans les enveloppes prédites. Ce qui conforte la validité du modèle.

Un deuxième outil diagnostique est l'analyse des npde (Normalised Prediction Distributions Errors) présentés en figure 3.8. L'interprétation de ces graphiques s'effectue en considérant que l'erreur de prédiction doit suivre une loi normale centrée réduite N(0,1). Les Q-QPlot, mais aussi la représentation des npde en fonction des prédictions (PRED) ou du temps (t), permet de s'assurer que les résidus sont répartis de façon équilibrée entre valeurs positives et négatives et selon une amplitude indépendante du niveau de PRED. Les résultats dans le cadre de gauche en A. concerne 324 observations pour 24 épidémies. Les courbes bleues sont bien dans les zones bleues, la courbe rouge est bien dans la zone rose, éventuellement, un seul petit bémol peut être remarqué après 13 semaines, ceci est dû à mon choix de ne prendre que les incidences concernant les périodes épidémiques, et donc quand t augmente, de moins en moins de points sont observés. J'aurais pu envisager l'option de prendre le même nombre de points pour chaque épidémie en gardant alors les incidences "négligeables" de début et de queues d'épidémies.

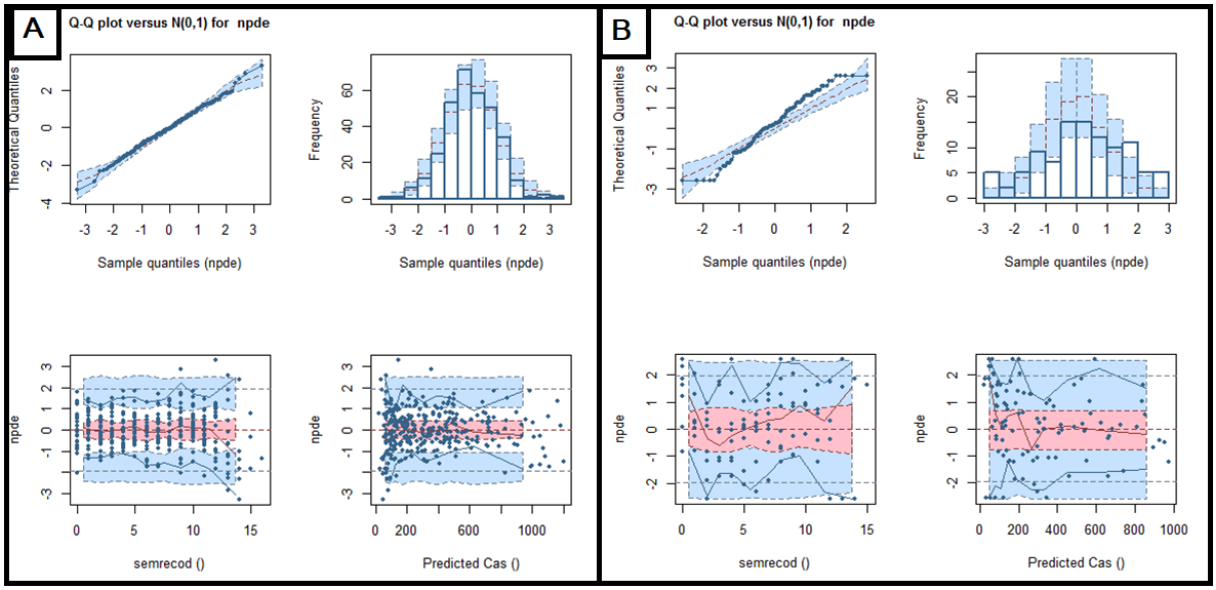


FIGURE 3.8 – Graphiques diagnostiques : les npde des 24 épidémies utilisées pour construire le modèles en A. et des 8 épidémies validantes en B.

En regardant les courbes épidémiques individuelles (figure 3.9), le modèle capte bien les rebonds des épidémies 6, 8, 14, 17 et 29.

En simulant sous le modèle les courbes épidémiques individuelles des 8 épidémies de validation, une variabilité colossale est alors observée (figure 3.10), mais qui inclut cependant l'épidémie observée.

En regardant les npde dans le cadre de droite en B. de la figure 3.8 qui concerne 101 observations pour 8 épidémies, nous pouvons noter que les zones bleues et roses se touchent, ce qui est logique avec seulement 8 sujets. De plus le grand nombre d'observations (12 observations par sujets en moyenne) pour si peu de sujets diminue l'intérêt de l'analyse de la distribution des npde. Notons donc juste la courbe

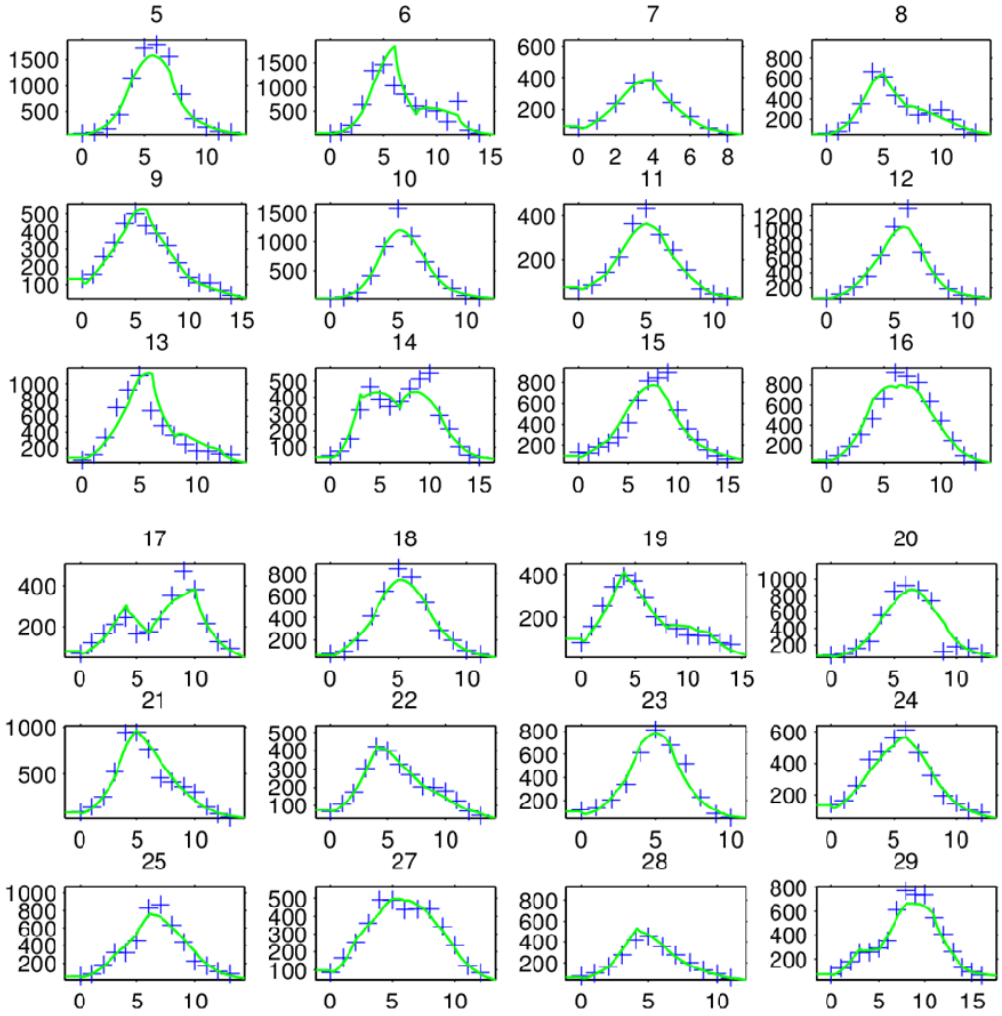


FIGURE 3.9 – Les 24 épidémies ayant servi à la construction du modèle

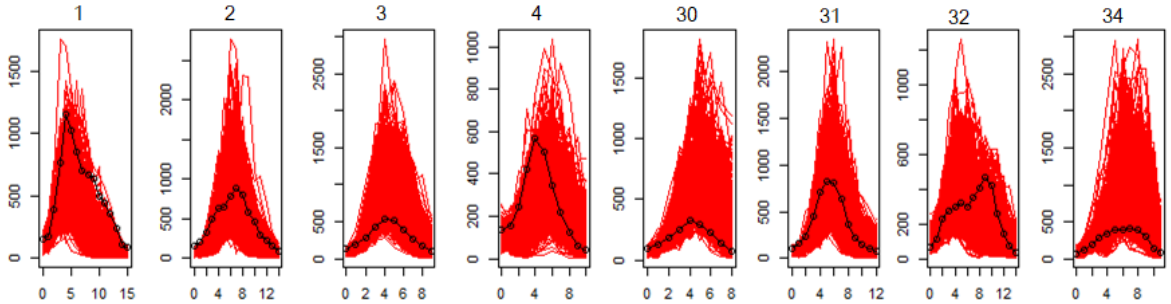


FIGURE 3.10 – Les 8 épidémies de validation

rouge qui sort de la zone rose quand t est grand ou proche de 0. Pour t grand, cela rejoint la remarque précédente sur les 24 épidémies et les t grands qui concernent peu d'épidémies. Pour la surestimation de l'incidence moyenne au temps 0, ce diagnostic est intéressant et là aussi, une autre option aurait été envisageable pour la modélisation. J'ai estimé un paramètre $I0_{pop}$ et une variabilité inter-individuelle puisque les observations démarrent seulement au dépassement du seuil épidémique. Peut-être que là aussi avoir le début "entier" de la courbe pourrait corriger ce problème.
Pour notre objectif explicatif des paramètres de population, ceci ne remet donc pas en cause le modèle.
Pour l'intérêt épidémiologique, l'analyse aurait dû prendre en compte une structure d'âge, j'ai obtenu les données du Réseau Sentinelles, j'ai complexifié mon modèle avec une matrice de contacts entre les classes d'âge, et malgré l'apport de données, le nombre de paramètres à estimer dans ce modèle très alourdi, a conduit à une modélisation qui n'était plus satisfaisante.
J'ai alors préféré mettre en suspend ce travail pour lancer une collaboration anglaise en revenant à la base de l'étude d'identifiabilité des modèles SIR. Quand cette partie, décrite dans le paragraphe suivant aura bien avancé, nous pourrons reprendre le travail précédent et envisager aussi un objectif prédictif avec la prédiction dynamique individuelle qui se développe en pharmacocinétique [107, 108]. Une collaboration se dessine avec Solène Desmée, maîtresse de conférences à Tours dans l'unité Inserm SPHERE, qui était en post-doc dans notre centre d'investigation clinique, et qui est spécialiste de la prédiction dynamique individuelle à partir de modèles non linéaires à effets mixtes. Elle utilise l'algorithme Hamiltonian Monte Carlo implémenté dans le logiciel Stan et les paramètres de population estimés dans l'étude comme a priori, la distribution a posteriori de la cinétique est calculée pour un nouvel individu statistique connaissant ses mesures jusqu'à un moment donné. Plusieurs questions peuvent alors être étudiées :

- A partir de quand les paramètres individuels deviennent stables ?
- Peut-on définir un indicateur de stabilité, qui permettrait de savoir prédire correctement à 3 semaines, le moment du pic et la taille de l'épidémie ?
- L'étude des épidémies saisonnières peut-elle permettre la prédiction de la pandémie de grippe H1N1 de 2009 (en fixant p_{S_0} à 100%-$p_{vaccination}$) ?

Diagnostics d'identifiabilité des paramètres des modèles SIR (grippe et covid19)

De nombreuses études manquent de condition préalable fondamentale au problème d'estimation des paramètres, à savoir l'identifiabilité structurelle de l'épidémie, c'est à dire la possibilité de déterminer de manière unique les paramètres du modèle à partir des données sur l'épidémie [109],[110]. Cela dépend aussi du type de données ajustées, c'est différent si nous avons des prévalences ou des incidences.

Identifiabilité structurelle et pratique L'identifiabilité structurelle est une caractéristique de la structure du modèle pour une sortie donnée (prévalence, incidence ou incidence cumulée) et elle repose sur l'hypothèse que le modèle est exempt d'erreurs et que la sortie est sans bruit. Mais, l'identifiabilité pratique dépend non seulement de la structure du modèle, mais aussi de la quantité et de la qualité des données ainsi que de l'algorithme d'optimisation numérique utilisé pour le problème d'estimation des paramètres. Si un modèle épidémiologique est structurellement non identifiable, alors il est clair qu'il est pratiquement non identifiable aussi. D'autre part, un modèle qui est structurellement identifiable peut être pratiquement non identifiable. Il est donc crucial d'examiner si des paramètres structurellement identifiables peuvent être estimés avec une précision correcte à partir de données bruitées en mettant en œuvre un algorithme d'optimisation. Dans cette étude, nous compléterons l'analyse de l'identifiabilité pratique pour le modèle de propagation d'épidémie [109] par l'optimisation bayésienne des paramètres (méthode ABC [111] - Approximate Bayesian Computation) et d'autres outils diagnostiques [112].

Modèles étudiés (SIR avec p_{S_0}, SIR avec k "R0", SAIR avec k "R0") Nous démarrons un travail de simulations pour envisager des recommandations pratiques et pédagogiques pour diagnostiquer facilement l'identifiabilité avant toute utilisation de modèles compartimentaux. La pandémie de coronavirus (COVID-19) a considérablement accru la sensibilisation du public et son appréciation de l'utilité des modèles dynamiques. Mais la diffusion de prévisions contradictoires des modèles a mis en évidence leurs limites. Un article récent [113] a passé en revue les différents modèles proposés dans la littérature les premiers mois de la pandémie, avec 36 structures de modèles et en évaluant leur capacité à fournir des informations fiables pour différentes situations. Cette étude approfondie montre par exemple que le taux de transmission est identifiable seulement pour 59/98.
Nous utiliserons trois modèles pour appliquer nos recommandations d'outils diagnostiques.

SIR avec paramètres supplémentaires (pourcentage de susceptibles et nombre d'infectieux au temps 0) S est le nombre de susceptibles au temps t, I le nombre d'infectieux au temps t, R le nombre de "retirés du système" au temps t, N la taille de population.

$$\begin{cases} \frac{dS}{dt} = -\frac{\gamma}{\theta p}\frac{SI}{N} \\ \frac{dI}{dt} = \frac{\gamma}{\theta p}\frac{SI}{N} - \frac{I}{\theta} \\ \frac{dR)}{dt} = \frac{I}{\theta} \end{cases}$$

avec γ le nombre de reproduction au temps 0, θ le temps de génération moyen, p le pourcentage de susceptibles au temps 0.
Avec la vaccination et les infections naturelles des années précédentes, p est souvent bien inférieur à 100%. Comme les valeurs proches de 0 ne sont pas informatives pour la dynamique de l'épidémie, nous avons besoin d'un autre paramètre. Plusieurs choix s'offre à nous : un décalage possible de l'injection des premiers infectieux, un décalage possible du moment où le nombre de reproduction est supérieur à 1, ou le nombre d'infectieux au temps 0.
Nous étudions ici le modèle SIR avec 4 paramètres : γ, θ, p et I au temps 0.

SIR avec k périodes avec ou sans mesures de confinement ou couvre-feu

$$\begin{cases} \frac{dS}{dt} = -\frac{\gamma_t}{\theta p_t}\frac{SI}{N} \\ \frac{dI}{dt} = \frac{\gamma_t}{\theta p_t}\frac{SI}{N} - \frac{I}{\theta} \\ \frac{dR}{dt} = \frac{I}{\theta} \end{cases}$$

avec γ_t le nombre de reproduction au temps t dépendant si la période est avec ou sans mesures de confinement ou couvre-feu $\gamma_t = \sum_{i=1}^{k} \gamma_i \mathbf{1}_{\{t \in \text{période} i\}}$ et p_t est le pourcentage de susceptibles au début de la période donnée t (soient k différentes valeurs de plus en plus petites et p est la première). Les périodes sont définies par les dates officielles de confinement et couvre-feu, et pourraient être décalées par un délai fixé avec la littérature ou une analyse de sensibilité.

SAIR avec k périodes avec ou sans mesures de confinement ou couvre-feu
Ici nous étudions le modèle comprenant les compartiments Susceptible, Asymptomatic, Infected et Recovered (SAIR modèle) [114].

$$\begin{cases} \frac{dS}{dt} = -\frac{\gamma_t}{\theta p_t}\frac{S(I+A)}{N} \\ \frac{dA}{dt} = \frac{\gamma_t}{\theta p_t}\frac{S(I+A)}{N} - \delta A - \frac{A}{\theta} \\ \frac{dI}{dt} = \delta A - \frac{I}{\theta} \\ \frac{dR}{dt} = \frac{A+I}{\theta} \end{cases}$$

Les asymptomatiques deviennent symptomatiques avec un taux effectif δ ou guérissent avec le même taux θ que les patients infectieux symptomatiques.

Application sur données réelles Ce dernier modèle a été appliqué à la 2ème vague de COVID19 au Royaume Uni et en France. Les données de covid19 sont publiques et faciles d'accès au travers du package R covid19.analytics pour de nombreux pays [115]. J'ai extrait les données quotidiennes françaises et britanniques à partir du 18/08/2020 sur 6 mois. Les données françaises quotidiennes présentaient quelques valeurs négatives, qui ne figuraient dans l'extraction (publique également) de Santé Publique France (https ://www.data.gouv.fr/fr/datasets/taux-dincidence-de-lepidemie-de-covid-19). C'est donc ces dernières que j'ai choisi d'utiliser pour la France. Par ailleurs, à cause des faibles dépistages le week-end, j'ai choisi une échelle de temps de 3 jours en sommant les incidences quotidiennes. Au Royaume Uni les deux périodes de confinement sur ces 6 mois furent : du 5 novembre au 2 décembre 2020 (effet z_1) et du 4 janvier à mars 2021 (effet z_3) En France, il y a eu un confinement du 30 octobre 2020 au 15 décembre 2020 (effet z_1), puis un couvre-feu à 21h (effet z_2), et de nouvelles mesures plus strictes du 16 janvier à début mars 2021 (effet z_3). Avec pS0 fixé à 85%, ce premier travail a abouti aux estimations suivantes :

$$\begin{cases} \frac{dS}{dt} = -\frac{R_0 \times z_{T_i}}{\theta p S T_i} \frac{S(I+A)}{N} \\ \frac{dA}{dt} = \frac{R_0 \times z_{T_i}}{\theta p S T_i} \frac{S(I+A)}{N} - \delta A - \frac{A}{\theta} \\ \frac{dI}{dt} = \delta A - \frac{I}{\theta} \\ \frac{dR}{dt} = \frac{A+I}{\theta} \end{cases}$$

	R0	z_1	z_2	z_3	θ	δ
France	1.47	0.521	0.840	0.745	2.47	0.34
Royaume-Uni	1.07	0.911	1.063	0.907	2.12	1.14

avec z_i l'effet des mises en place ou arrêt de mesures de confinement ou couvre-feu sur le nombre de reproduction de base et $z_{T_i} = \sum_{i=1}^{k} z_i \mathbf{1}_{\{t \in \text{période } i\}}$

Une estimation de δ à 0.34 s'interprète comme 8.82 jours pour passer de l'état asymptomatique (A) à symptomatique (I), et θ à 2.47 comme 7.41 jours pour passer de l'état symptomatique (I) à guéri (R). L'estimation des z_1, z_2, et z_3 vont dans le sens d'une efficacité des mesures de confinement et couvre-feu, et l'ajustement du modèle semble correct au vu des courbes de la figure 3.11.

Le travail évoqué précédemment de diagnostic d'identifiabilité des paramètres se verra ici naturellement appliqué. C'est une collaboration anglaise avec Dave Woods, et Samuel Jackson qui a été reçu 2 semaines à l'IRMAR sur financements per diem pour travailler sur ce projet avec moi en septembre 2022.

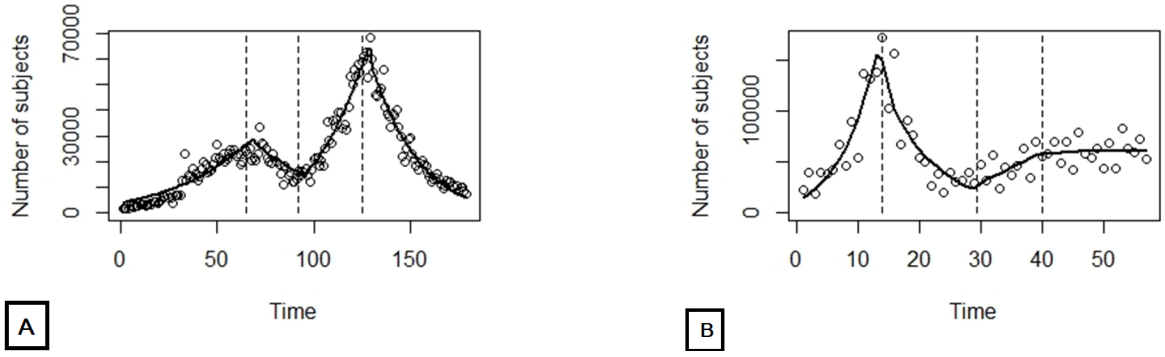


FIGURE 3.11 – Modélisation de la 2ème vague de covid19. A. au Royaume Uni, B. en France.

Ma contribution à la propagation d'épidémie en termes de modélisation compartimentale SIR

La construction de modèles compartimentaux se base sur le développement de nouveaux systèmes d'équations différentielles régissant des transferts de personnes dans les compartiments d'état relatif à la maladie (notre article [1]), ou de concentration de molécules d'intérêt dans le sang par exemple (notre article [6]), nécessite de pouvoir traduire mathématiquement un phénomène biologique à partir de la littérature et/ou de discussions avec les cliniciens, mais aussi de contrôler chaque étape de construction pour obtenir le meilleur modèle aux estimateurs non biaisés qui pourront être cliniquement interprétés.
La construction de modèles non linéaires à effets mixtes nécessite également souvent ces compétences mathématiques à coupler aux connaissances médicales, même si le modèle structural n'est pas un modèle compartimental, comme dans le modèle sigmoïde à effets mixtes que j'ai construit pour étudier la croissance tumorale [5]. La particularité des modèles conjoints est très souvent un développement spécifique d'un modèle complexe à poser et formuler mathématiquement [6]. Mon projet de modélisation compartimentale à effets mixtes des épidémies de grippe est un autre exemple de formalisation mathématique innovante que j'ai pu imaginer par décloisonnement disciplinaire. La construction de tous ces modèles complexes nécessite des vérifications d'hypothèses statistiques parfois à définir spécifiquement à l'aide de certains outils diagnostiques émergeant dans la littérature. Ma collaboration avec Dave Woods et Samuel Jackson (Southampton et Durham) porte sur le développement d'un nouveau critère statistique de diagnostic de non identifiabilité des paramètres. J'ai initié et je porte cette étude théorique et expérimentale qui utilise simulations et données réelles. J'ai également participé au développement d'un package R (saemix) (article [90]). Deux autres papiers dont je suis co-auteure ont été évoqués dans ce chapitre car portant sur les données utilisées pour la modélisation (articles [8, 2]).

Chapitre 4

Modélisation des données de survenue d'événement(s) dans un cadre de grande dimension

Ce chapitre qui résonnera avec le premier chapitre par rapport à la thématique de la grande dimension et des méthodes d'apprentissage qui y ont été présentées, s'attache à la spécificité des données de survie. Il repose sur

- quatre articles de statistique appliquée sur la dichotomisation de variables explicatives quantitatives dans des modèles de survie en cancérologie,
- un travail en cours basé sur des simulations de données de survie en grande dimension dont la première partie a été publiée dans les actes des journées de la statistique de la SFDS 2022, en collaboration avec deux chercheuses de l'IRMAR (Magalie Fromont et Valérie Garès), et
- le premier article de thèse de Juliette Murris que je co-encadre qui comprend une revue de la littérature sur les méthodes d'analyse d'événements récurrents en grande dimension et une étude de simulation.

Dans ce chapitre, nous adoptons les notations classiques des données de survie. Nous considérons donc un échantillon de n individus. Pour chaque individu i, T_i est le temps de survie pas toujours observé, C_i est le temps de censure, et $\Delta_i = \mathbf{1}_{T_i \leq C_i}$ indique s'il y a eu absence de censure ou pas. La variable observée est $T_i^O = T_i \wedge C_i = \min(T_i, C_i)$. Les observations de ces variables aléatoires sont respectivement notées t_i, c_i, δ_i, t_i^O. Pour illustrer ces notations spécifiques aux données de survie, la figure 4.1 montre un exemple avec 4 individus : les deux premiers avec leur durée de survie t_i qui n'est pas observée car ils sont décédés après la date de point (date de fin de collecte des données). Le troisième pour qui t_3 est connue, et le quatrième qui est perdu de vue (avec une croix rouge sur le graphique de gauche). Les durées t_i^O représentées sur le graphique de droite sont donc les durées de vie observées correspondant à la valeur minimale entre la durée de survie t_i et la durée de censure c_i, complétée de la croix rouge qui indique que la durée est censurée et donc que δ_i vaut 0.

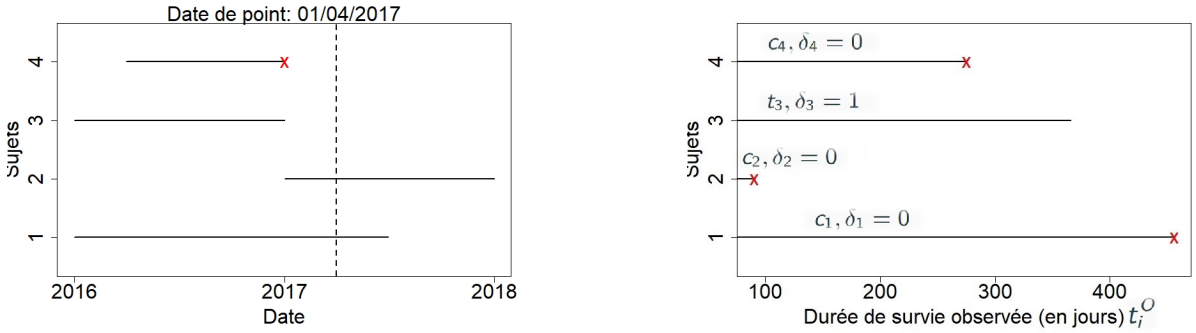


FIGURE 4.1 – Illustration des censures dans les données de survie

Notations

T_i temps de survie
C_i temps de censure
$T_i^O = T_i \wedge C_i$
$\Delta_i = \mathbf{1}_{T_i \leq C_i}$
X_i vecteur des covariables
$\longrightarrow$ Les variables observées sont
$\mathscr{D}_1 = (T_1^O, \Delta_1, X_1), \ldots, \mathscr{D}_n = (T_n^O, \Delta_n, X_n)$
$\longrightarrow$ L'ensemble des données est noté
$\mathscr{D} = \{\mathscr{D}_1, \ldots, \mathscr{D}_n\}$

4.1 Quelques bases pour le traitement des données de survie

4.1.1 Le modèle de Cox

À nos durées de survie observées et nos indicateurs de censure, $(T_1^O, \Delta_1), (T_2^O, \Delta_2), \ldots, (T_n^O, \Delta_n)$, pour tout $1 \leq i \leq n$, on ajoute $X_i = (X_i^1, \ldots, X_i^p)$ un vecteur de p covariables. Le modèle de Cox est basé sur une expression de la fonction de risque instantanné de décès au temps t qui considère des risques proportionnels, en multipliant un risque de base par l'exponentielle du prédicteur linéaire :

$$\lambda(t, X_i) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j X^j\right).$$

On introduit la log-vraisemblance dite partielle définie par :

$$\ell((t_i, \delta_i, x_i)_{i=1,\ldots,n}, \beta) = \lim_{h \to 0} \frac{\mathbb{P}\Big(T_i \in [t, t+h[\mid T_i \geq t\Big)}{h} = \sum_{i=1}^{n} \delta_i \Big[F_\beta(x_i) - \ln \sum_{i'=1}^{n} \Big(\mathbf{1}_{t_i^O \leq t_{i'}^O} \exp(F_\beta(x_{i'}))\Big)\Big] \; \forall t \in \mathbb{R},$$

avec $F_\beta(x_i) = \Sigma_{j=1}^{p} \beta_j x_i^j$. Cette vraisemblance partielle n'est pas une vraisemblance dans le sens statistique du terme parce qu'elle ne prend en compte que les t_i^O telles que $\delta_i = 1$, mais elle se comporte comme

telle. Ainsi, une théorie asymptotique similaire a été développée qui justifie son utilisation pour estimer et tester les coefficients de régression β.

Évoquons également ici le test du log-rank qui sera cité à plusieurs reprises. Il a pour but de comparer deux fonctions de survie définies par deux groupes de sujets différents (en imaginant une covariable X à 2 modalités). Ce test fonctionne aussi pour plus de deux groupes, mais mes applications dans ce mémoire ne portent que sur deux groupes. Il a été proposé par Mantel en 1966 [116] et nommé test du log-rank un peu plus tard par Peto [117]. Il est basé sur un test du χ^2 stratifié (test de Mantel-Haenszel), où des tableaux de contingence 2×2 sont construits à chaque temps de décès $t(1) < t(2) < \ldots < t(K)$ avec les 4 effectifs décédés versus vivants en fonction des deux groupes. Notons d_{1k} le nombre de décédés dans le groupe 1 du k^e tableau de contingence, d_k le nombre total de décédés dans les deux groupes, n_{1k} le nombre d'individus du premier groupe, n_{2k} le nombre du deuxième groupe, et $n_k = n_{1k} + n_{2k}$.

t(k)	Décédés	Vivants	Total
Groupe 1	d_{1k}	$n_{1k} - d_{1k}$	n_{1k}
Groupe 2	$d_k - d_{1k}$	$n_{2k} - (d_k - d_{1k})$	n_{2k}
	d_k	$n_k - d_k$	n_k

Sous l'hypothèse nulle d'égalité des lois des temps de survies des deux groupes, D_{1k} suit une loi hypergéométrique, donc $\mathbb{E}(D_{1k}) = E_k = \frac{n_{1k}d_k}{n_k}$, $Var(D_{1k}) = V_k = \frac{n_{1k}n_{2k}d_k(n_k-d_k)}{n_k^2(n_k-1)}$, et la statistique de test du log-rank définie par $U_{log-Rank} = \frac{\left(\sum_{k=1}^K (D_{1k}-E_k)\right)^2}{\sum_{k=1}^K V_k}$ suit asympotiquement la loi χ^2_{1ddl}.

4.1.2 La performance du modèle

La capacité prédictive d'un modèle de survie peut être résumée à l'aide d'extensions de la proportion de variation expliquée par le modèle comme :

- le R^2, couramment utilisé pour les modèles de réponse continue, ou
- des extensions de sensibilité et de spécificité, qui sont couramment utilisées pour les modèles de réponse binaire.

L'aire sous la courbe ROC temps-dépendante

En 2005, dans l'article [118] sont proposés de nouveaux indicateurs de précision dépendants du temps basés sur des versions temporelles de la sensibilité et de la spécificité calculées à partir du modèle de régression standard de Cox. Pour les données de survie, il existe plusieurs extensions possibles de la sensibilité et de la spécificité. Un temps de survie peut être considéré comme un résultat binaire variable dans le temps en se concentrant sur la représentation du processus de comptage $N_i^*(t) = \mathbf{1}_{\{T_i<t\}}$. Plusieurs définitions sont alors possibles pour la sensibilité dépendante du temps, nous choisissons celle dite des cas *incident* où $T_i = t$ (i.e $dN_i^*(t) = 1$, avec $dN_i^*(t) = N_i^*(t+dt) - N_i^*(t)$).

$$\text{sensibilité}(c,t) = \mathbb{P}(M_i > c | T_i = t) = \mathbb{P}(M_i > c | dN_i^*(t) = 1),$$

$$\text{spécificité}(c,t) = \mathbb{P}(M_i \leq c | T_i > t) = \mathbb{P}(M_i \leq c | N_i^*(t) = 0),$$

où M_i est un score de risque ou de mortalité prédit par le modèle pour l'individu i (pour le modèle de Cox on utilise souvent directement $\sum_{j=1}^p \hat{\beta}_i x_i^j$, où $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ est le vecteur des coefficients de régression estimés par maximisation de la log vraisemblance partielle $\ell((t_i, \delta_i, x_i)_{i=1,\ldots,n}, \beta)$.
En utilisant cette approche, un sujet peut jouer le rôle d'un contrôle pour un temps précoce, $t < T_i$, et jouer le rôle de cas quand $t = T_i$. Ce statut dynamique est parallèle aux multiples contributions qu'un sujet

peut faire à la fonction de vraisemblance partielle. Ici, la sensibilité mesure la fraction attendue de sujets avec un score supérieur à c parmi la sous-population de personnes qui meurent au moment t, alors que la spécificité mesure la fraction de sujets avec un score inférieur ou égal à c parmi ceux qui survivent au-delà du temps t. La sensibilité *incident* et la spécificité *dynamique* sont définies en dichotomisant le groupe des personnes à risque au temps t en un groupe des *cas* (ceux qui seront observés pour mourir) et un groupe *contrôle* (ceux qui seront observés pour survivre). La courbe ROC (Receiver Operating Characteristics) à chaque temps correspond à la courbe paramétrée définie par $((1 - \text{spécificité}(c,t)), \text{sensibilité}(c,t)), \forall c \in [0,1]$. Définissons ROC_t comme la fonction représentée par la courbe ROC au temps t. L'aire sous la courbe ROC s'écrit donc :

$$\text{AUC}(t) = \int_0^1 \text{ROC}_t(p)dp,$$

Les auteurs de l'article [118] montrent que $\text{AUC}(t) = \mathbb{P}(M_j > M_k | T_j = t, T_k > t)$ et construisent un indice de concordance pour une durée de suivi τ donné :

$$C^\tau = \int_0^\tau \text{AUC}(t) w^\tau(t) dt,$$

avec $w^\tau(t) = \frac{2f(t)S(t)}{W^\tau}$, où $f(t)$ est la densité marginale de T_i, $S(t)$ est sa fonction de survie en t soit $\mathbb{P}(T_i > t)$, $W^\tau = \int_0^\tau 2f(t)S(t)dt = 1 - S^2(\tau)$ qui permettent de mettre à l'échelle les poids $w^\tau(t)$ de sorte que leur intégrale sur $(0,\tau)$ soit égale à 1. L'interprétation de C^τ est une légère modification de la concordance originale définie par $\mathbb{P}[M_j > M_k | T_j < T_k, T_j < \tau]$. Donc C^τ est la probabilité que les prédictions soient concordantes avec les observations pour une paire aléatoire de sujets, sachant que le plus petit temps d'événement a lieu entre 0 et τ.

Le C de Harrell

Une autre manière d'aborder la question est de comparer toutes les paires d'individus deux à deux, au lieu de décompter les bien classés ou mal classés parmi les personnes à risque à chaque temps. On s'intéresse à la probabilité de concordance entre les scores prédits par le modèle et les durées de survie $\mathbb{P}[M_i < M_{i'} | T_i > T_{i'}]$, basé sur le principe de concordance quand le modèle donne un score de risque ou de mortalité d'un individu i plus faible que pour l'individu i', sachant que sa durée de survie est plus grande que celle de i', que l'on estime seulement quand les comparaisons entre les deux individus sont possibles, c'est à dire quand la durée la plus petite est inférieure aux durées de censure des deux. L'estimateur considéré ici est appelé le C de Harrell ou C-index de Harrell, défini par :

$$\text{C}_{\text{Index}} = \frac{\sum_{i=1}^n \sum_{i'=1, i \neq i'}^n \mathbf{1}_{\{T_i^O > T_{i'}^O\}} . \mathbf{1}_{\{M_{i'} > M_i\}} . \delta_{i'}}{\sum_{i=1}^n \sum_{i'=1, i \neq i'}^n \mathbf{1}_{\{T_i^O > T_{i'}^O\}} \delta_{i'}}.$$

qui calcule la proportion de couples (i,j), où prédictions et observations sont concordantes au sens décrit ci-dessus.

4.1.3 Application à la recherche de seuils de biomarqueurs en survie

L'étude ECOM est une étude en cancérologie initiée par le Centre Eugène Marquis. Les glioblastomes multiformes (GBM) correspondent à des tumeurs gliales cérébrales primitives de haut grade de malignité. Ce sont des tumeurs intra-crâniennes, très exceptionnellement métastatiques. Ce cancer atteint

sa fréquence maximale vers 64 ans mais peut survenir à tout âge. On évaluait à l'époque de l'étude à environ 2000 le nombre de nouveaux cas de GBM opérés chaque année en France [119]. Le pronostic en était particulièrement sombre [1] : la survie des patients était de 6 mois après traitement chirurgical seul, et prolongée seulement de quelques mois par la radiothérapie. La chimiothérapie adjuvante n'avait pas fait preuve d'efficacité réelle, mais venait de rentrer dans les standards de traitement depuis la mise en évidence d'un bénéfice en terme de survie de l'association du témozolomide (TMZ) à la radiothérapie [120]. La plupart des patients nouvellement pris en charge et opérables étaient traités par chirurgie, suivie d'un traitement concomitant TMZ/radiothérapie et de six cycles adjuvants de chimiothérapie à base de TMZ. Les données de la littérature indiquaient cependant que certains patients étaient non répondeurs au TMZ. Cette chimiorésistance était suspectée d'être principalement liée à la présence, au sein des cellules tumorales, d'une enzyme de réparation O6-methylguanine-DNAmethyltransferase (MGMT). Le lien entre sensibilité au TMZ et présence de l'enzyme MGMT (ARN, protéine et activité enzymatique) a formellement été montré sur des lignées cellulaires de gliomes malins [121]. Il existait d'autre part une littérature abondante qui montrait que la détermination de la MGMT au niveau de la tumeur était corrélée à la survie chez les patients atteints d'un GBM et traités par chimiothérapie alkylante. Dans la plupart des études publiées, c'était la méthylation au niveau de la région promotrice du gène MGMT qui était étudiée [122], celle-ci entraînant une inactivation du gène et donc une diminution de l'activité enzymatique. Il était également possible d'étudier l'expression du gène au niveau ARN ou encore de rechercher directement l'expression de la protéine par des techniques d'immunohistochimie, mais avant de proposer dans les laboratoires la détermination du statut MGMT pour les patients atteints d'un glioblastome, il était primordial de déterminer sur une cohorte homogène de patients, la technique d'analyse qui réunirait les meilleures propriétés quant à la valeur prédictive, le coût, la praticabilité et la reproductibilité.

C'est dans ce cadre qu'a été proposée la mise en place d'un essai multicentrique français en 3 étapes. La première étape a consisté à tester sur une cohorte rétrospective différentes techniques d'analyse, chacune étant centralisée dans le laboratoire en ayant déjà la maîtrise. A la fin de cette étape, deux techniques ont été retenues et diffusées dans les différents centres participants avec notamment la mise en place d'un système qualité comprenant la distribution de Contrôles qui permettront d'habiliter les différents centres. La dernière étape a consisté en l'analyse prospective, dans les différents centres, de la MGMT sur les patients traités selon le protocole standard comprenant du TMZ. Cette démarche de validation était fondamentale, alors que le test MGMT commençait à apparaître dans les arbres décisionnels de prise en charge des gliomes de haut grade. Sans une telle étude, il était à craindre que les différents laboratoires ne développent des techniques « propres », validées sur des petites cohortes d'échantillons, techniques qui pourraient ensuite être utilisées pour le rendu de résultat sans véritable validation à grande échelle. Ce projet a été fortement soutenu par l'AnOcef (Association des neuro-oncologues d'expression française).

Je décrirai ici surtout les résultats de recherche de seuils, ceux-ci ont été définis comme les seuils optimisant C^τ pour τ fixé à 18 mois (la durée de suivie) obtenus avec un modèle de Cox de la survie globale ajustée sur l'âge et le score de Karnofsky. Le C^τ et l'indice C de Harrell ont été calculés pour chaque modèle. Dans le premier article [3], la recherche de seuils s'est faite pour 5 sites CpG situés au niveau de l'exon 1 du gène MGMT analysés par la technique de pyroséquençage, et pour la technique d'immunohistochimie. Cette dernière est une technique simple de lecture de lames, très dépendante de l'anatomopathologiste (la concordance des résultats de deux lecteurs a également été étudiée par le test de Bland-Altman). Le graphique 4.2 montre à gauche C^τ en ordonnée en fonction de c. Le seuil optimisé est : $\text{argmax}_c\, C^\tau$, que l'on prendra à 11 car compris entre 10 et 12. La figure de droite montre les courbes de survie observées et prédites par notre modèle de Cox construit avec l'âge et le score de Karnofsky et

1. Malgré des progrès indéniables, c'est toujours le cas aujourd'hui, d'après le Dr Jean Ménard, oncologue à l'Hôpital Saint-Louis à Paris "le pronostic du glioblastome reste sombre, avec une survie à 5 ans avoisinant les 5%, et un taux médian de survie entre 14 et 18 mois en 2019".

Technique	No. (%)	Median OS, months	HR	*Pvalue*	Median PFS, months	HR	*Pvalue*
PYRCpG2			0.31	$2.4e+5$		0.40	$1.2e+4$
> 11	38 (38)	26.2			15.3		
≤ 11	61 (62)	15.8			9.0		

TABLE 4.1 – Extrait de la Table 2 de l'article [3]. Statut MGMT selon les différentes techniques et leur significativité concernant la survie (OS) ou la (PFS) en analyse univariée.

CpG2 dichotomisé par le seuil de 11.
Les résultats présentés dans l'article figurent sur la table 4.1.

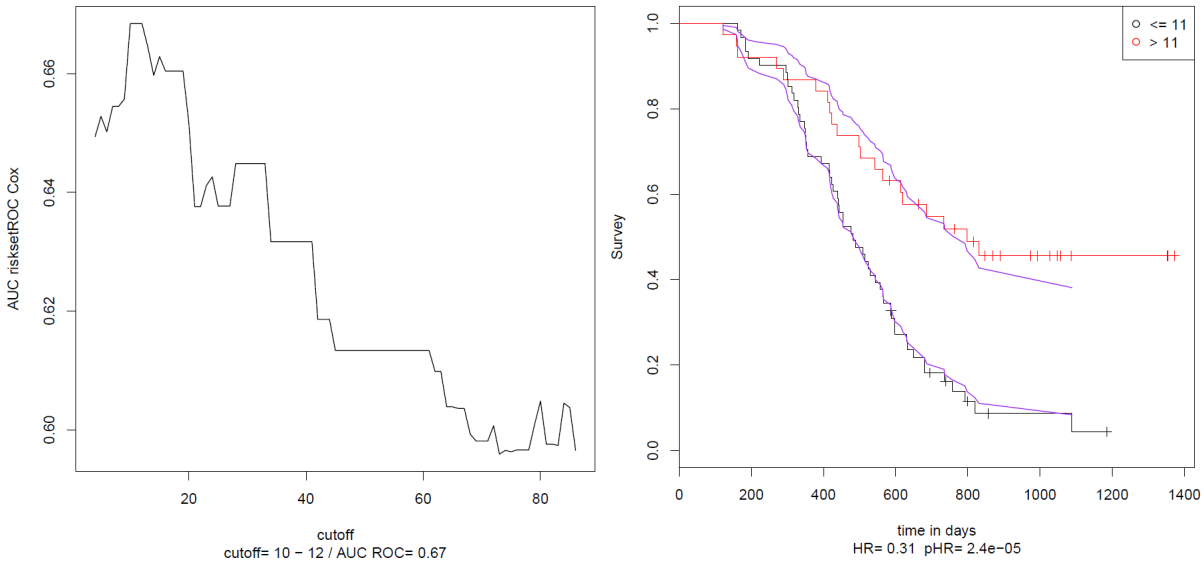


FIGURE 4.2 – Recherche de seuil optimal pour le CpG2. A gauche l'optimisation du critère C^τ. A droite les 2 courbes de survie obtenues avec ce seuil.

Dans cet article, nous avons donc comparé les performances analytiques et les valeurs prédictives de 5 techniques dont deux quantitatives (pyroséquençage de 5 sites CpG et immunohistochimie) et trois qualitatives (MS-PCR, methylation-sensitive high-resolution melting (MS-HRM) et MethyLight) pour l'analyse MGMT chez 100 patients GBM traités par temozolomide. Le pyroséquençage est la meilleure méthode parmi les 5 techniques testées dans cette étude. Cet article a été cité 102 fois (requête pubmed entre le 18/10/12 et le 24/07/22).

Les trois articles qui ont suivi ([4], [7], [10]) ont utilisé les mêmes méthodes de recherche de seuils. Au moment de la publication de notre papier sortait cet article [123], cité à ce jour par 540 papiers, qui venait de développer Cutoff Finder, un ensemble de méthodes d'optimisation et de visualisation pour la détermination de seuils de biomarqueurs. Dans ce papier était écrit qu'il n'existait pas de méthode ou de logiciel standard, alors que les études cliniques ont si souvent besoin de traduire une variable continue dans une décision clinique en déterminant un point de coupure et de stratifier les patients en deux groupes nécessitant chacun un traitement différent.

A cette période sortait également le papier [124] qui comparait par simulation 2 méthodes :

— **Minimisation d'une fonction de coût basée sur les courbes ROC dépendantes du temps utilisant l'estimateur de Kaplan-Meier**. Cette méthode est basée sur la courbe

ROC temps-dépendante, décrite précédemment, avec une estimation non paramétrique [125] par Kaplan-Meier de la probabilité de survie au temps t conditionnelle à $X > c$. Nous avons utilisé ici le prédicteur linéaire du modèle de Cox.

— **Minimisation d'une fonction de coût basée sur les courbes ROC dépendantes du temps utilisant l'estimateur d'Akritas (plus proches voisins)** : Heagerty et al. [125] ont également proposé d'estimer la spécificité en utilisant l'estimateur des plus proches voisins initialement proposé par Akritas [126]. Cet estimateur assure une courbe ROC monotone contrairement à l'approche Kaplan-Meier. De plus, l'estimateur Kaplan-Meier sera biaisé si la censure dépend du marqueur, alors que celui d'Akritas sera robuste dans ce cas.

4.2 Survie en grande dimension

L'utilisation des mégadonnées dans la recherche et la pratique en santé publique explose aujourd'hui, et les questions sous-jacentes d'analyse de survie en grande dimension deviennent incontournables. La figure 4.3 montre l'évolution de l'utilisation de méthodes spécifiques dans les articles de recherche médicale ou appliquée à la médecine (données issues d'une requête Pubmed que j'ai précisée en bas de page [1]).

Les technologies modernes permettent de générer des données sur des milliers de variables ou d'observations, selon la génomique, la radiomique, les bases de données médico-administratives, la surveillance des maladies par des dispositifs médicaux intelligents, etc. Alors que les données massives décrivent un grand nombre d'observations, les données de grande dimension sont caractérisées lorsque le nombre de variables étudiées p est supérieur au nombre d'individus n. C'est précisément le contexte de grande dimension qui sera considéré ici. Il se peut que les modèles statistiques standards ne soient plus applicables dans ce cas, car ils ont tendance à faire face à des problèmes de convergence, d'instabilité, et une signification non pertinente sur le plan clinique des variables peut survenir. Pour aider à résoudre des problèmes de grande dimension, de nombreuses méthodes d'apprentissage automatique ont émergé. Sur la base d'une

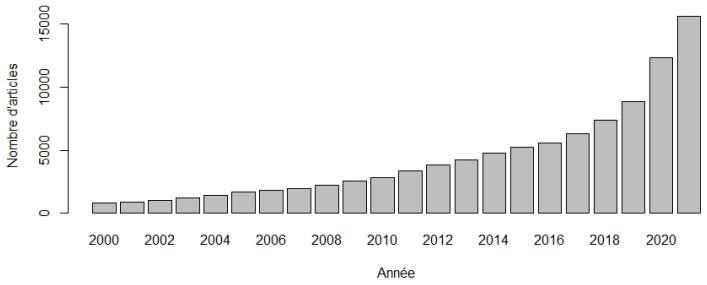


FIGURE 4.3 – Nombre d'articles référencés sur Pubmed qui utilisent des méthodes statistiques et/ou d'apprentissage pour la prédiction du risque d'une réponse de type temps d'événement dans des données médicales ou de santé. Les critères choisis sont ceux de l'article [22] sans la restriction "simulation" et élargis jusqu'à fin 2021.

revue de la littérature des méthodes statistiques couramment utilisées et des techniques d'apprentissage automatique développées pour l'analyse de survie, une taxonomie détaillée des méthodes existantes a été

1. (english[LANGUAGE])AND("2000/01/01"[Date - Publication] : "2021/12/31"[Date - Publication])AND ((("machine learning"[Title/Abstract] OR "ai"[Title/Abstract] OR "ml"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "neural network*"[Title/Abstract] OR "ann*"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "random forest*"[Title/Abstract] OR "random survival forest*"[Title/Abstract] OR "bayesian learning"[Title/Abstract] OR "bayesian network*"[Title/Abstract] OR "support vector*"[Title/Abstract] OR "svm"[Title/Abstract] OR "svms"[Title/Abstract]) AND ("survival"[Title/Abstract] OR "hazard"[Title/Abstract] OR "risk"[Title/Abstract] OR "prognos*"[Title/Abstract] OR "time to event"[Title/Abstract] OR "censor*"[Title/Abstract] OR "cox"[Title/Abstract] OR "kaplan*"[Title/Abstract] OR "spline*"[Title/Abstract])))

proposée en 2019 dans l'article [127] et a été reprise et modifiée encore plus récemment dans l'article [22] (figure 4.4).

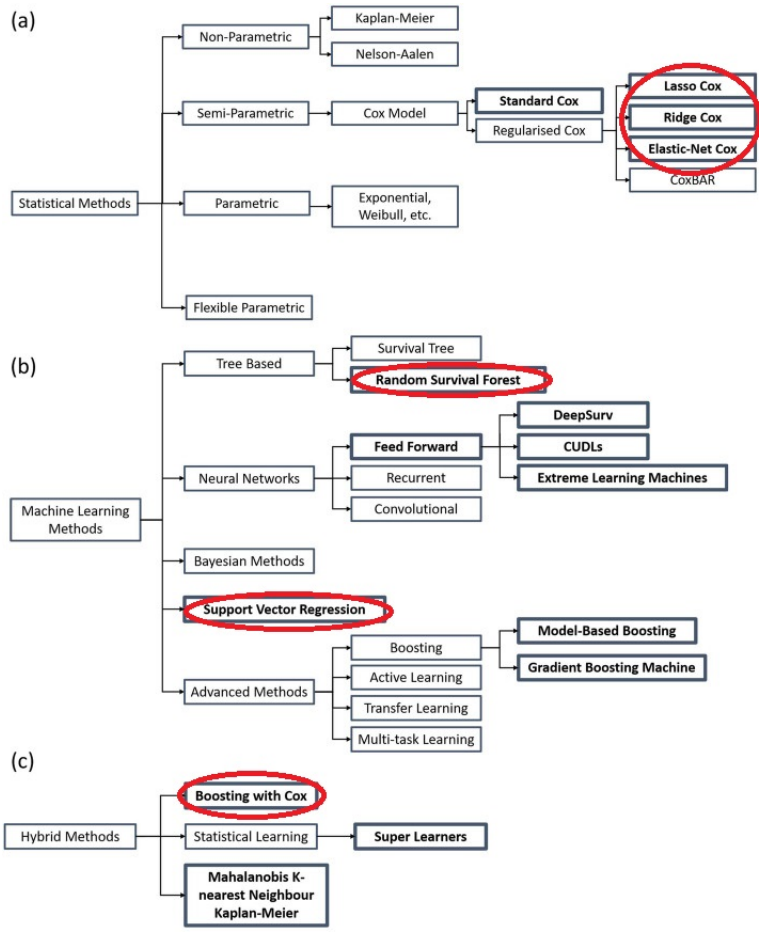


FIGURE 4.4 – Figure 1 de l'article [22] Taxonomie des méthodes pour la modélisation pronostique comme définie dans cette revue de la littérature, adapté de la taxonomie de Wang et al. [127]. Les méthodes étaient classées comme statistiques (a), machine learning (b), ou méthodes hybrides (c). Sont repérées en gras si elles sont étudiées dans les 10 articles inclues dans cette revue. Sont entourées en rouge les méthodes présentes dans mon mémoire d'HDR.

Dans les études pronostiques, nous avons deux objectifs : la prédiction (c'est-à-dire la précision de l'estimation du risque) et la détection des facteurs de risque (c'est-à-dire l'interprétation des résultats en fonction de l'importance des covariables), même si le débat entre ces deux objectifs existe depuis longtemps et a justement été ravivé depuis le développement des méthodes d'apprentissage supervisé, comme discuté dans l'introduction. Bien que la détection des facteurs de risque soit considérée nécessaire pour développer un meilleur diagnostic et des stratégies de traitement optimales [128], ces deux objectifs sont rarement étudiés simultanément.

Tous les algorithmes d'apprentissage impliquent un plus ou moins grand nombre d'hyperparamètres (ou "paramètres de réglage"). La quantité de gain de performance qui peut être obtenue en optimisant les hyperparamètres par rapport aux valeurs par défaut définit la "tunabilité". Il a été montré que la méthode

d'optimisation des hyperparamètres et le moment où elle se fait peut engendrer des biais sur les petits échantillons, par exemple un fort taux de bien classés d'une variable binaire alors que les covariables étaient simulées sans effet [129].

L'objectif de nos travaux est, à terme, de comparer les performances des méthodes pour analyser des données de survie de grande dimension, donc souvent sur petits échantillons avec beaucoup de covariables. Pour la méthode Cox Boost [130] en particulier, nous étudierons ici ses performances en termes de prédiction et de discrimination des variables pronostiques, mais aussi sa tunabilité en fonction des tailles d'échantillon et du taux de covariables informatives.

4.2.1 Trois méthodes de survie adaptées à la grande dimension

Cox Boost

L'ajustement des modèles de survie par la vraisemblance partielle de Cox est la méthode la plus largement utilisée pour la régression des risques proportionnels linéaires. Le boosting s'adapte aussi aux modèles de régression non linéaires et aux modèles de Cox en construisant une famille d'estimateurs qui sont ensuite agrégés.
Deux méthodes existent : celle basée sur des arbres que nous n'avons pas utilisée, et celle de descente de gradient avec recherche directionnelle par composante qui autorise de s'arrêter en laissant un grand nombre de paramètres à 0.

On introduit la log-vraisemblance partielle de Cox pénalisée de type "ridge" comme dans l'article [131] pour la j^e composante et l'étape k de l'algorithme défini pour $\gamma \in \mathbb{R}$ par :

$$\ell^{\lambda}_{n,j,k}(\gamma) = \sum_{i=1}^{n} \delta_i \left[x_i \cdot \widehat{\beta}_{k-1} + x_i^j \gamma - \ln \left(\sum_{i'=1}^{n} \mathbf{1}_{t^O_{i'} \leq t^O_i} e^{x_{i'} \cdot \widehat{\beta}_{k-1} + x^j_{i'} \gamma} \right) \right] - \lambda \gamma^2, \text{ où}$$

- $(t_i^O, \delta_i, x_i), i \in \{1, \ldots, n\}$ sont les données observées.
- $\widehat{\beta}_{k-1} = (\widehat{\beta}_{(k-1,1)}, \ldots, \widehat{\beta}_{(k-1,p)}) \in \mathbb{R}^p$ obtenu à la $(k-1)^e$ étape de l'algorithme.
- $x_i \cdot \widehat{\beta}_{k-1} = \sum_{j=1}^{p} x_i^j \widehat{\beta}_{(k-1,j)}$.

La log-vraisemblance partielle non pénalisée correspondante sera notée $\ell^0_{n,j,k}$.

Description de l'algorithme Cox Boost component-wise [130, 132] :

Cette description de la méthode est surtout basée sur l'article [130] dans une version plus générale, où l'on considère des blocs de composantes que l'on s'autorise à pénaliser ou non, selon les étapes du boosting.

Ici, on considère une seule composante par bloc et on pénalise par le même constante systématiquement.

Algorithme CoxBoost Component-wise

1. Initialiser : $\widehat{\beta}_0 =^t (0, \ldots, 0) \in \mathbb{R}^p$
2. Pour $k = 1$ à K_{stop}, répéter les étapes suivantes :
   (a) Pour tout $j \in \{1, \ldots, p\}$ calculer
   $\widehat{\gamma}_{k,j} = \frac{\frac{\partial \ell^0_{n,j,k}(\gamma)}{\partial \gamma}|_{\gamma=0}}{-\frac{\partial^2 \ell^0_{n,j,k}(\gamma)}{\partial \gamma^2}|_{\gamma=0}+\lambda}$;
   (b) Déterminer $j_k^* = \text{argmax}_{1 \leq j \leq p} \ell^\lambda_{n,j,k}(\widehat{\gamma}_{k,j})$
   (donne la composante à mettre à jour) ;
   (c) Mettre à jour $\widehat{\beta}_k$:
   $\widehat{\beta}_{k,j} = \begin{cases} \widehat{\beta}_{k-1,j_k^*} + \widehat{\gamma}_{k,j_k^*} \text{ si } j = j_k^* \\ \widehat{\beta}_{k-1,j} \text{ si } j \neq j_k^* \end{cases}$
   (une seule des composantes est mise à jour).

L'étape 2.(a) vient d'une méthode de Newton Raphson (adaptée au cas pénalisé) pour la minimisation de $-\ell^0_{n,j,k}(\gamma)$ en γ.

Random Survival Forest (RSF)

Nous avons vu le principe des forêts aléatoires dans le chapitre 2.

L'arbre de survie est similaire à l'arbre de décision qui est construit par division récursive des nœuds de l'arbre. Un nœud d'un arbre de survie est considéré " pur " si tous les patients du nœud survivent pendant une durée identique.

La statistique du test du log-rank est la mesure de dissimilarité la plus couramment utilisée pour estimer la différence de survie entre deux groupes. Pour chaque nœud, il s'agit d'examiner toutes les répartitions possibles sur chaque covariable, puis de sélectionner la meilleure répartition, celle qui maximise la différence de survie entre deux nœuds enfants [133].

L'association du principe des forêts aléatoires et de l'arbre de survie donne donc la méthode Random Survival Forest (dite RSF) [134, 135].

Algorithme RSF

Entrée

• Les données observées $\mathscr{D} = \{\mathscr{D}_1, \ldots, \mathscr{D}_n\}$ avec $\mathscr{D}_1 = (T_1^O, \Delta_1, X_1), \ldots, \mathscr{D}_n = (T_n^O, \Delta_n, X_n)$
- $T_i^O = T_i \wedge C_i$ où T_i temps de survie et C_i temps de censure,
- $\Delta_i = \mathbf{1}_{T_i \leq C_i}$,
- X_i vecteur des covariables.
• x le vecteur des covariables pour lequel on souhaite faire la prédiction.
• Les hyperparamètres :
- m le nombre de covariables sélectionnées à chaque nœud (Maximal number of features),
- B le nombre d'itérations (Number of survival trees),
- Critère à maximiser pour définir les nœuds engendrant des groupes à différence de survie la plus grande (Splitting rule) qui peut prendre les valeurs "Log-rank", "C-index", ou "maximally selected rank statistics".

Pour b variant de 1 à B :
1. Tirer un échantillon bootstrap $\mathscr{D}^{*b} = \left\{\mathscr{D}_1^{*b}, \ldots, \mathscr{D}_n^{*b}\right\}$ de $\mathscr{D}^*$ (en moyenne asymptotiquement 63.2% d'individus différents composeront nos échantillons bootstrap de n individus, conduisant à une taille moyenne des échantillons OOB (Out Of Bag) de 36.8% de n).
2. Construire un arbre de survie de taille maximale (dont chaque nœud terminal contient au moins un décès) sur $\mathscr{D}^{*b}$.
A chaque nœud de l'arbre :
- sélectionner m covariables candidates aléatoirement,
- diviser le nœud en deux nœuds enfants en choisissant la covariable candidate et le seuil associé qui maximisent la différence de survie entre les deux nœuds enfants selon le critère choisi en hyperparamètre "Splitting rule".

Sorties ou prédictions

• La fonction de risque cumulée (CHF) estimée par la forêt aléatoire, définie par :

$$\mathrm{CHF}^*(t|x) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{CHF}^{*b}(t|x) \quad \forall t > 0,$$

où $\mathrm{CHF}^{*b}(.|x)$ est l'estimateur de Nelson-Aalen [136] de la fonction de risque cumulé calculé sur les observations du nœud terminal ou de la feuille à laquelle x appartient dans l'arbre de survie construit sur le b^e échantillon bootstrap défini par :

$$\mathrm{CHF}^{*b}(t|x) = \sum_{t_{l,f_x^{*b}} < t} \frac{d_{l,f_x^{*b}}}{r_{l,f_x^{*b}}}, \text{où}$$

f_x^{*b} désigne la feuille à laquelle x appartient, $t_{l,f_x^{*b}}$ est le l^e temps d'évènement observé dans la feuille f_x^{*b}, $d_{l,f_x^{*b}}$ est le nombre d'évènements observés à $t_{l,f_x^{*b}}$ dans la feuille f_x^{*b}, $r_{l,f_x^{*b}}$ est le nombre d'individus à risque à $t_{l,f_x^{*b}}$ dans la feuille f_x^{*b}.
• $\mathrm{CHF}^{*\mathrm{OOB}}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b}\mathrm{CHF}^{*b}(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}$, avec $I_{i,b}$ fixé à 1 si i est un cas OOB pour l'échantillon bootstrap b ; 0 sinon.

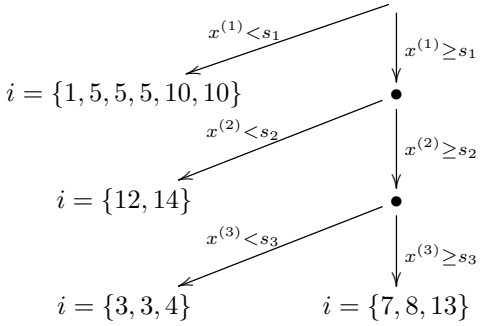
Calcul du C de Harrell (avec un score de risque ou de mortalité M_i spécifique) :
Le score qui est utilisé ici est appelé mortalité globale OOB défini par :

$$M_i = \sum_{l=1}^{L} \text{CHF}^{*\text{OOB}}(t_l^O | x_i)$$

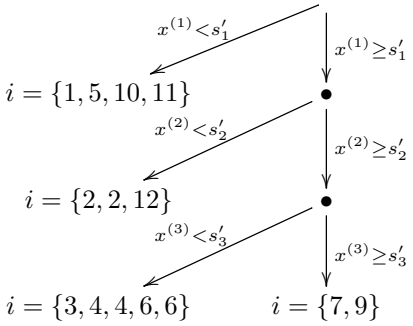
où $\{t_1^O, \ldots, t_l^O\}$ sont les temps d'évènements observés sur l'ensemble des individus de départ.

Exemple : $n = 14; B = 4$. Pour simplifier l'exemple, le nombre de nœuds (4) est le même pour tous les arbres et les conditions aux nœuds sont proches (ici variables identiques, seuls les seuils changent s_k, s'_k, s''_k, s'''_k pour k allant de 1 à 4, sans conséquence sur les regroupements d'individus), mais le principe restera le même : chaque vecteur x "tombera" dans une feuille.

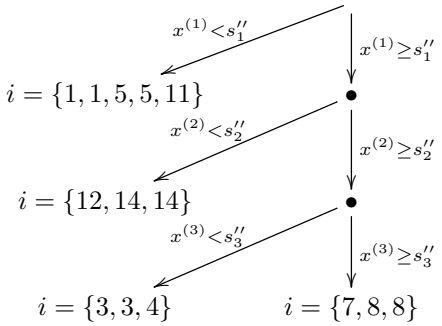
1er tirage bootstrap remis dans l'ordre :
$i = \{1, 3, 3, 4, 5, 5, 5, 7, 8, 10, 10, 12, 13, 14\}$
Donc $OOB = \{2, 6, 9, 11\}$



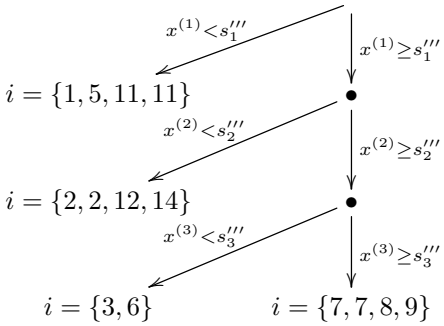
2eme tirage bootstrap remis dans l'ordre :
$i = \{1, 2, 2, 3, 4, 4, 5, 6, 6, 7, 9, 10, 11, 12\}$
Donc $OOB = \{8, 13, 14\}$



3eme tirage bootstrap remis dans l'ordre :
$i = \{1, 1, 3, 3, 4, 5, 5, 7, 8, 8, 12, 14, 14\}$
Donc $OOB = \{2, 6, 9, 10, 11, 13\}$



4eme tirage bootstrap remis dans l'ordre :
$i = \{1, 2, 2, 3, 5, 6, 7, 7, 8, 9, 11, 11, 12, 14\}$
Donc $OOB = \{4, 10, 13\}$



Par exemple pour l'individu $i = 10$ avec le vecteur des covariables x_{10}, la CHF^{*1} est calculée sur les individus $\{1, 5, 5, 5, 10, 10\}$, la CHF^{*2} est calculée sur les individus $\{1, 5, 10, 11\}$, etc. Et $\text{CHF}^{*\text{OOB}}(t|x_{10})$ est calculée sur les arbres 3 et 4.

Survival Support Vector Machine (SSVM)

Nous avons vu dans le chapitre 2 le principe des SVM et de leur extension à une variable réponse continue (SVR). L'analyse de survie peut être vue comme un problème de régression et il suffit alors d'adapter la SVR à la thématique des données censurées [137]. Mais elle peut également être vue comme un problème de classement [138] (nous pensons d'ailleurs immédiatement au C de Harrell dans ce cas), ce qui a conduit à des extensions des SVM de rang (RankSVM) [139, 140]. Enfin, Van Belle et al. [141] ont proposé une solution hybride entre l'approche de classement (rangs) et de régression.
L'article [142] décrit les deux méthodes comme suit.

SSVM$_{\text{regression}}$. Contrairement à un modèle basé sur le classement, un modèle de régression peut prédire le moment exact d'un événement (i.e. le temps de survie). En analyse de survie, les algorithmes d'apprentissage d'un tel modèle doivent tenir compte des observations censurées. Pour les patients censurés à droite - ceux qui n'ont pas connu d'événement - aucune information sur l'exactitude des temps de survie prédits au-delà du moment de la censure n'est disponible. Une erreur valide ne peut donc être calculée que pour les patients qui ont connu un événement pendant la période d'étude, ou si le temps de survie prédit est précoce, c'est-à-dire avant le moment de la censure. L'article [141] a révélé que les modèles de survie basés sur la ξ-insensible SVR fonctionnaient aussi bien si la zone insensible est fixée à zéro. Par conséquent, la fonction objective de régression est basée sur un problème de moindres carrés ordinaires avec la pénalité $\mathbb{L}^2$ et la considération supplémentaire de la censure à droite. On cherche alors $\text{argmin}_{w,b} f_{Regr.}(w,b)$, où

$$f_{Regr.}(w,b) = \frac{1}{2}{}^t ww + C\sum_{i=1}^{n}(\zeta_{w,b}(t_i^O, x_i, \delta_i))^2$$

$$\zeta_{w,b}(t_i^O, x_i, \delta_i) = \begin{cases} \max(0, t_i^O - {}^t wx_i - b) & \text{si } \delta_i = 0, \\ y_i - {}^t wx_i - b & \text{si } \delta_i = 1, \end{cases}$$

où $w \in \mathbb{R}^p$ sont les coefficients de poids et $C > 0$ est le paramètre de régularisation.
SSVM$_{\text{Rang}}$. Avec les rangs, en analyse de survie, l'objectif est de retrouver l'ordre correct des individus en fonction de leur durée de survie. Cependant, toutes les comparaisons par paires ne sont pas significatives en présence de censure à droite. Soit l'ensemble $\mathscr{P} = \left\{(i,i')|t_i^O > t_{i'}^O, \delta_{i'} = 1\right\}_{i,i'=1,\ldots,n}$ définit les paires d'échantillons comparables qui peuvent être utilisées pour l'apprentissage. La fonction objective de l'article [143] peut être modifiée pour tenir compte en plus de la censure à droite pendant l'apprentissage. Sera donc minimisée :

$$f(w) = \frac{1}{2}{}^t ww + C\sum_{i,j\in\mathscr{P}} \max(0, 1 - {}^t wx_i - {}^t wx_j))^2.$$

Un nouveau jeu de données x_{new} peut être rangé selon les temps de survie prédits d'après les valeurs de ${}^t wx_{\text{new}}$.

Le score utilisé pour le calcul du C de Harrell est alors $M_i = -f(x_i)$.

La fonction FastKernelSurvivalSVM() implémentée sous Python est issue de l'article [144] qui propose un algorithme d'apprentissage efficace basé sur une SSVM$_{\text{Rang}}$ à noyau.
L'idée principale pour obtenir une fonction de prédiction non linéaire est que le problème d'optimisation précédent est reformulé pour trouver une fonction $f : \mathscr{X} \longrightarrow \mathbb{R}$ dans un espace de représentation $\mathscr{H}_k$ associé à un noyau $k : \mathscr{X} \times \mathscr{X} \longrightarrow \mathbb{R}$, donnée par :

$$\underset{f\in\mathscr{H}_k}{\text{argmin}} \frac{1}{2}||f||^2_{\mathscr{H}_k} + C\sum_{i,j\in\mathscr{P}} \max(0, 1 - (f(x_i) - f(x_j)))^2.$$

L'optimisation de Newton tronquée et des arbres statistiques d'ordre sont utilisés dans l'article [144] pour réduire de manière significative les coûts de calcul par rapport aux algorithmes similaires précédents.

4.2.2 Critères de performance, d'ajustement des hyperparamètres et d'importance des variables

L'objectif visé dans notre étude étant le classement plutôt que la prédiction du temps de survie, le critère de performance que nous avons considéré est la probabilité de concordance estimée par le C de Harrell, dont on rappelle la définition ci-dessous :

$$\mathrm{C}_{\mathrm{Index}} = \frac{\sum_{i=1}^{n}\sum_{i'=1,i\neq i'}^{n} \mathbf{1}_{\left\{T_i^O > T_{i'}^O\right\}}.\mathbf{1}_{\{M_{i'}>M_i\}}.\delta_{i'}}{\sum_{i=1}^{n}\sum_{i'=1,i\neq i'}^{n} \mathbf{1}_{\left\{T_i^O > T_{i'}^O\right\}}\delta_{i'}},$$

où M_i est un score de risque ou de mortalité prédit par le modèle pour le i^e individu.

Estimation de la probabilité de concordance. Pour l'ajustement des paramètres, afin d'éviter les phénomènes de surapprentissage, une estimation par des méthodes de validation croisée (ou Out of Bag dans le cas des RSF) est nécessaire.
Sur de petits échantillons, les méthodes de validation croisée (VC) à K blocs sont alors privilégiées. On considère une partition $\{\mathscr{I}_1, \ldots \mathscr{I}_K\}$ de $\{1, \ldots, n\}$. Pour chaque $k = 1, \ldots, K$, le C de Harrell est calculé sur l'échantillon test $\left\{(x_i, t_i^O, \delta_i), i \in \mathscr{I}_k\right\}$ avec un score M_i^k construit sur la base de l'échantillon d'apprentissage $\left\{(x_i, t_i^O, \delta_i), i \in \{1, \ldots, n\} \setminus \mathscr{I}_k\right\}$. Il est noté $\mathrm{C}_{\mathrm{Index}}^k$.
On considère ensuite le C de Harrell moyen sur les K blocs $\mathrm{C}_{\mathrm{Index}}^{\mathrm{VC}} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{C}_{\mathrm{Index}}^k$.
Plus le nombre de blocs augmente, plus la taille de l'échantillon d'apprentissage est grande, plus celle de l'échantillon test est petite. A noter qu'une validation croisée avec un seul individu par bloc ("leave-one-out") ne peut pas être réalisée de cette manière.

Méthodes algorithmiques pour le choix des hyperparamètres. On sait que les algorithmes de prédiction (méthodes) dépendent d'hyperparamètres $h \in \mathscr{H}$ à ajuster. Soit $\mathscr{G} = \{g_h, h \in \mathscr{H}\}$, la famille des algorithmes g dépendant de l'hyperparamètre h. L'objectif est de sélectionner l'algorithme g_h qui minimise le C de Harrell $\mathrm{C}_{\mathrm{Index}}^{\mathrm{VC}}$. Nous avons choisi deux approches d'optimisation des hyperparamètres.

a. La méthode HalvingGridSearch. Cette méthode [145] consiste en l'évaluation à partir d'un sous-ensemble des données plutôt que sur toutes les données. A chaque étape du processus, les combinaisons d'hyperparamètres les plus performantes sont évaluées et la quantité de données utilisées augmente en se concentrant sur les meilleures combinaisons. On dit qu'il s'agit d'halving successifs.
Ce processus est répété jusqu'à la dernière itération, où il ne reste que quelques candidats (en fonction de la taille de la grille d'hyperparamètres). Le meilleur candidat est celui qui a la meilleure performance à la dernière itération. L'algorithme par halving successifs permet un gain de temps considérable par rapport à une méthode GridSearch simple qui permet de parcourir une grille d'hyperparamètres de manière exhaustive (vs. RandomSearch).

b. La méthode Optuna. Cette méthode [146] est un système d'optimisation automatique par recherche dynamique des hyperparamètres. A chaque itération, cet algorithme permet de tirer au hasard une combinaison d'hyperparamètres en se basant sur les performances de la meilleure combinaison des itérations précédentes, et stoppe les combinaisons les moins prometteuses.
Optuna est organisé en plusieurs modules :

— **Study**

  - Gère la fonction objectif pour la meilleure combinaison d'hyperparamètres
  - Contrôle la méthode d'optimisation et le nombre de tests (les trials) à effectuer
- **Trial**
  - Contrôle la fonction objectif et l'ensemble des hyperparamètres
  - Envoie les informations dans Storage
- **Sampler** Exécute le processus d'échantillonnage bayésien des hyperparamètres à partir d'un estimateur de Tree-Parzen (TPE)

Algorithme Optuna

1. Sélection aléatoire d'un sous-ensemble d'hyperparamètres et tri en fonction de la performance
2. Création de deux groupes d'hyperparamètres et estimation des densités de Parzen
3. Identification des hyperparamètres avec l'amélioration attendue (expected improvement) la plus élevée
4. Evaluation et tri des hyperparamètres de l'étape 3. et nouvelle création de deux groupes

Ce processus est répété suivant le nombre de "trials" fixé.

Critère d'importance des variables basé sur le C de Harrell. Pour mesurer l'importance de la variable j, on considère la partition $\{\mathscr{I}_1, \dots \mathscr{I}_K\}$. Pour chaque $k = 1, \dots, K$, on considère $\left\{x_i^j, i \in \mathscr{I}_k\right\}$ l'ensemble des vecteurs de covariables pour les individus de $\mathscr{I}_k$ où les valeurs de la je variable ont été perturbées aléatoirement entre ces individus. Le score M_i^k a été défini précédemment. On calcule ensuite le C de Harrell sur l'échantillon ainsi "perturbé" $\left\{(x_i^j, t_i^O, \delta_i), i \in \mathscr{I}_k\right\}$ que l'on note $\mathrm{C}_{\mathrm{Index}}^{j,k}$. La mesure d'importance de la variable j est donnée par $\mathrm{Imp}(j) = \frac{1}{K}\sum_{k=1}^{K}(\mathrm{C}_{\mathrm{Index}}^{k} - \mathrm{C}_{\mathrm{Index}}^{j,k})$.

Évaluation de la méthode. Pour les M jeux de données, nous avons estimé le C de Harrell de la méthode CoxBoost par la validation croisée à 2 et 5 blocs. Nous avons mis en œuvre trois méthodes de choix des hyperparamètres (sans optimisation donc avec les valeurs par défaut, et les deux méthodes précédemment citées : HalvingGridSearch et Optuna).

Sur des jeux de données simulés, nous faisons varier le taux appelé *sparse rate* ou taux de variables actives noté $sr = \frac{p'}{p}$, nous connaissons le vrai statut actif ou passif des p covariables dont p' sont actives, nous notons Y la variable aléatoire associée. Soit $Y^j = 1$ si la variable X^j est active (et a un effet sur la durée de vie) et $Y^j = 0$ sinon. Pour $s \in [0, 1]$ on définit :

$$\widehat{x}(s) = \frac{\sum_{j=1}^{p} \mathbf{1}_{\{\mathrm{Imp}(j) > s\}}(1 - Y^j)}{\sum_{j=1}^{p}(1 - Y^j)} \text{ et } \widehat{y}(s) = \frac{\sum_{j=1}^{p} \mathbf{1}_{\{\mathrm{Imp}(j) \geq s\}} Y^j}{\sum_{j=1}^{p} Y^j}, \quad \forall s \in [0, 1].$$

Nous appelons $\mathrm{AUC}^{\mathrm{Imp}}$ l'aire sous la courbe paramétrée définie par $(\widehat{x}(s), \widehat{y}(s)), \forall s \in [0, 1]$. Si $\mathrm{AUC}^{\mathrm{Imp}}$ est proche de 1, nous conclurons que la méthode de prédiction aura plutôt utilisé les variables actives sur ce jeu de données.

En notant $j_1, \dots, j_p$ les indices des covariables tels que $\mathrm{Imp}_{j_1} \geq \dots \geq \mathrm{Imp}_{j_p}$ nous pouvons en outre calculer $\sum_{k=1}^{p'} \mathbf{1}_{\left\{Y^{j_k}=1\right\}}$ le nombre de variables actives dans les p' variables dont la mesure d'importance est la plus grande, avec $p' = p.sr$ le nombre de variables simulées actives.

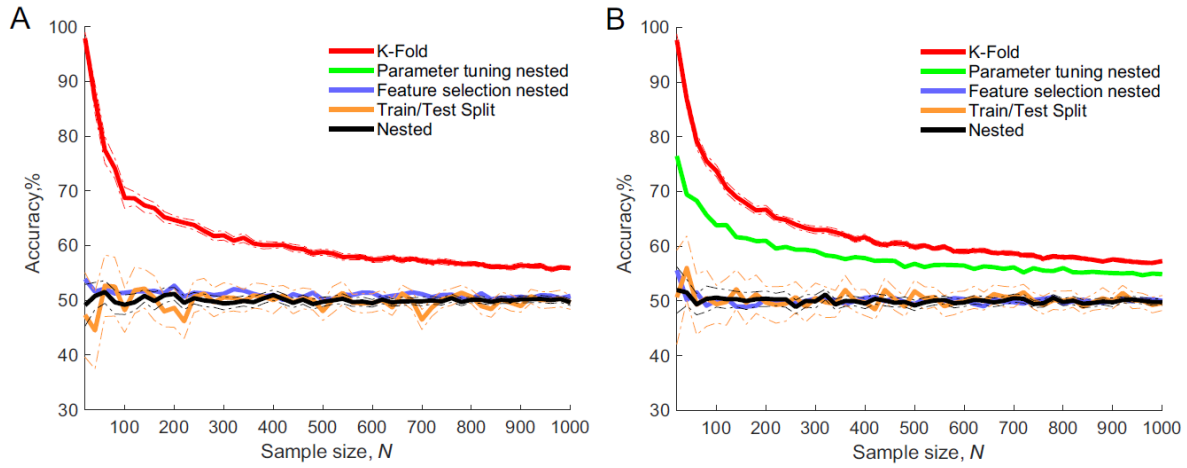


Fig 3. Gaussian noise classification accuracy distributions with different validation approaches. K-Fold, Nested, Train/Test Split and two types of partially nested validation methods used. Thick lines show mean validation accuracy and dash-dot lines show 95% confidence intervals for 50 runs. **A**: SVM-RFE feature selection and SVM classification. **B**: *t*-test feature selection and logistic regression classification.

https://doi.org/10.1371/journal.pone.0224365.g003

FIGURE 4.5 – Figure 2 de l'article [129]. Biais de la performance induit par les méthodes de validation en fonction de la taille d'échantillon.

4.2.3 Simulations

Rappelons le contexte de nos travaux très classique dans la recherche médicale en radiomique ou génomique, avec en conséquences un nombre de sujets très inférieur au nombre de covariables, et donc des spécificités pour la validation des modèles et des spécificités pour l'optimisation des hyperparamètres des méthodes. La figure 4.5 publiée en 2019 montre, pour des données binaires, des biais des métriques selon la taille d'échantillon et la méthode de validation. Par exemple, le K-fold classique en rouge, où l'optimisation des hyperparamètres se fait sur l'échantillon entier donne un *accuracy* complètement biaisé puisque les simulations ont été faites sans covariable informative et que l'accuracy devrait donc être à 50%.

Génération de données et schémas de simulation

Génération de données. Rappelons nos notations : nous considérons un échantillon de n individus, pour chaque individu i, soient X_i le vecteur de ses p covariables, T_i son temps de survie, C_i son temps de censure. Nous observons : $T_i^O = \min(T_i, C_i)$ et $\Delta_i = \mathbf{1}_{T_i \leq C_i}$. Soit $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ où $\mu = (a, \ldots, a)$ et Σ ayant une structure de corrélation autorégressive telle que $\text{corr}(X_i, X_j) = \rho^{|i-j|}$ ($\rho = 0.7$) avec $1 \leq i, j \leq p$. Basé sur un modèle de Cox avec composantes linéaires, nous associons les coefficients $\beta = (\beta_1, \ldots, \beta_p)$ aux p covariables. p' coefficients sont fixés à b (indicés $j_1, \ldots, j_{p'}$) et $p - p'$ sont fixés à 0. Le taux appelé *sparse rate* ou taux de variables actives est noté $sr = \frac{p'}{p}$. Prenons T un temps d'événement suivant une distribution de Weibull $\mathscr{W}(\lambda_x, \nu)$ où $\lambda_x = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x^j\right) = \alpha \exp\left(\sum_{j=1}^{p} \beta_j x^j\right)$ avec $\alpha = \exp(\beta_0)$, nous avons

$$\text{Med}(T|X = x) = \left(\frac{\log(2)}{\lambda_x}\right)^{1/\nu}$$

Si nous fixons m_0 la médiane de survie de base désirée (pour ${}^t x = (0, 0, \ldots, 0)$ et m_1 la médiane de survie

pour une augmentation de une unité d'une covariable active $x-(x-1)=(1,0,\ldots,0)-(0,0,\ldots,0)=(1,0,\ldots,0)$, alors ${}^t\beta(x-(x-1))=b$. Posons $m(x)=\text{Med}(T|X=x)$. Nous avons

$$
\begin{aligned}
m(x)=\left(\frac{\log(2)}{\lambda_x}\right)^{1/\nu}=\left(\frac{\log(2)}{\alpha\exp({}^t\beta x)}\right)^{1/\nu} \quad &\Longleftrightarrow \quad \alpha^{1/\nu}=\frac{1}{m(x)}\left(\frac{\log(2)}{\exp({}^t\beta x)}\right)^{1/\nu} \\
&\Longleftrightarrow \quad \alpha=\frac{1}{m(x)^\nu}\left(\frac{\log(2)}{\exp({}^t\beta x)}\right) \\
&\underset{\text{pour x=(0,0,...,0)}}{\Longleftrightarrow} \quad \boldsymbol{\alpha=\frac{\log(2)}{m_0^\nu}} \\
&\qquad\qquad \text{et } m_0=\left(\frac{\log(2)}{\alpha}\right)^{1/\nu} \\
&\underset{\text{pour x=(1,0,...,0)}}{\Longleftrightarrow} \quad m_1=\left(\frac{\log(2)}{\alpha\exp(b)}\right)^{1/\nu}
\end{aligned}
$$

Donc à partir du ratio $\frac{m_0}{m_1}=\exp(b)^{1/\nu}$, est déduit $\boldsymbol{b=\log\left(\left(\frac{m_0}{m_1}\right)^\nu\right)}$.

Avec la distribution de Weibull, nous avons aussi $F_{T|X=x}(t)=1-\exp(-\lambda_x t^\nu)$
Posons U qui suit une distribution uniforme $\mathscr{U}(0,1)$:

$$
\begin{aligned}
\exp(-\lambda_x T^\nu)=1-U \quad &\Longleftrightarrow \quad \lambda_x T^\nu=-\log(1-U) \\
&\Longleftrightarrow \quad T=\left(\frac{-\log(1-U)}{\lambda_x}\right)^{1/\nu}\sim\left(\frac{-\log(U)}{\lambda_x}\right)^{1/\nu}
\end{aligned}
$$

Nous simulerons donc T de cette manière, par la méthode d'inversion de la fonction de répartition, à partir de tirages pseudo-aléatoires qui seront des simulations de la loi uniforme.
Nous supposons que la censure C suit une distribution uniforme de paramètre θ, $C\sim\mathscr{U}([0,\theta])$ et est indépendante de T conditionnellement à X. Soit $\tau=\mathbb{P}(T<C)=\mathbb{P}(\Delta=1)$ le taux de censure.

Pour simuler les données de survie avec des taux de censure prédéfinis pour les modèles à risques proportionnels, nous avons utilisé la méthode de Wan [147]. Nous avons adapté les calculs à notre choix de loi normale pour les covariables X, de loi de Weibull pour T et de loi uniforme pour C, pour obtenir θ qui permettra de fixer les taux de censure à la valeur souhaitée même si les scénarios font varier le nombre de covariables actives ou le paramètre ν de la loi de Weibull.

Nous avons besoin de la distribution de $\lambda_X=\exp(\beta_0+\sum_{j=1}^p\beta_j X^j)$.

$$X\sim\mathscr{N}_p(\mu,\Sigma)$$

$$\sum_{j=1}^p\beta_j X^j=(\beta_1,\ldots,\beta_p)\begin{pmatrix}X^1\\X^2\\\ldots\\X^p\end{pmatrix}=UX$$

donc

$$\sum_{j=1}^p\beta_j X^j\sim\mathscr{N}_p(U(\mu),U\Sigma{}^tU)=\mathscr{N}_p\left(\sum_{j=1}^p\beta_j\mu_j,(\beta_1,\ldots,\beta_p)\Sigma\begin{pmatrix}\beta_1\\\beta_2\\\ldots\\\beta_p\end{pmatrix}\right)$$

et

$$\log(\lambda_i) = \log(\alpha) + \sum_{j=1}^{p} \beta_j X^j \sim \mathcal{N}_p \left(\log(\alpha) + \sum_{j=1}^{p} \beta_j \mu_j, (\beta_1, \ldots, \beta_p)\Sigma \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{pmatrix} \right)$$

la fonction de densité de λ_i est donnée par

$$\begin{aligned} f_{\lambda_i}(u) &= \frac{1}{u\sqrt{2\pi}\sqrt{{}^t\beta\Sigma\beta}} \exp\left(-\frac{(\log(u) - \log(\alpha) - \sum_{j=1}^{p} \beta_j \mu_j)^2}{2^t\beta\Sigma\beta} \right) \\ &= \frac{1}{u\sqrt{2\pi}\sqrt{{}^t\beta\Sigma\beta}} \exp\left(-\frac{(\log(u) - \log(\alpha) - mba)^2}{2^t\beta\Sigma\beta} \right) \end{aligned}$$

$$\begin{aligned} \mathbb{P}(\delta = 1|\lambda_x, \nu, \theta) = \int_0^{+\infty} f_C(u) F_{T|X=x}(u) du &= \int_0^{\theta} \frac{1}{\theta} \exp(-\lambda_x u^\nu) du \\ &= \int_0^{\lambda_x \theta^\nu} \frac{1}{\lambda_x^{1/\nu} \theta \nu} v^{\frac{1}{\nu} - 1} \exp(-v)\, dv \\ &= \frac{1}{\lambda_x^{1/\nu} \theta \nu} \Gamma\left(\frac{1}{\nu}, \lambda_x \theta^\nu \right) \end{aligned}$$

avec $v = \lambda_x u^\nu$ et $u = \left(\frac{v}{\lambda_x}\right)^{1/\nu}$ et $dv = \lambda_x \nu u^{\nu-1} du = \lambda_x^{1/\nu} \nu v^{1-1/\nu} du$ et $\Gamma(\cdot, \cdot)$ est une fonction gamma incomplète inférieure.

On obtient θ en résolvant $\gamma(\theta)$=0 avec

$$\begin{aligned} \gamma(\theta) &= \int_0^{+\infty} \mathbb{P}(\delta = 1|u, \nu, \theta) f_{\lambda_i}(u) du - \tau \\ &= \int_0^{+\infty} \frac{1}{\theta \nu u^{1/\nu}} \Gamma\left(\frac{1}{\nu}, u\theta^\nu \right) \frac{1}{u\sqrt{2\pi}\sqrt{{}^t\beta\Sigma\beta}} \exp\left(-\frac{(\log(u) - \log(\alpha) - \sum_{j=1}^{p} \beta_j \mu_j)^2}{2^t\beta\Sigma\beta} \right) du - \tau \\ &= \int_0^{+\infty} \frac{1}{\theta \nu u^{1/\nu}} \Gamma\left(\frac{1}{\nu}, u\theta^\nu \right) \frac{1}{u\sqrt{2\pi}\sqrt{{}^t\beta\Sigma\beta}} \exp\left(-\frac{(\log(u) - \log(\alpha) - mba)^2}{2^t\beta\Sigma\beta} \right) du - \tau \end{aligned}$$

Nous utilisons un algorithme d'intégration numérique pour calculer l'intégrale à l'intérieur de l'algorithme qui trouve les racines de $\gamma(\theta)$. Une fois θ calculé, nous simulons C_i indépendant de T_i selon une loi uniforme sur $[0, \theta]$. Nous construisons alors les données observées $T_i^O = \min(T_i, C_i)$ and $\delta_i = \mathbf{1}_{T_i \leq C_i}$.

Pour le cas particulier où le *sparse rate sr* est à 0 donc $p' = 0$, $\beta = (0, \ldots, 0)$ alors $\lambda_x = \alpha, \forall x$. Nous avons toujours $T \sim \mathscr{W}(\alpha, \nu)$ et la censure C une uniforme ($\mathscr{U}[0, \theta]$). Et nous avons ici $\tau = \mathbb{P}(\Delta = 1)$ que l'on veut égal à $\mathbb{P}(C \leq T)$ en calculant le paramètre θ de la loi de C par la résolution de $\gamma(\theta) = 0$.

$$\gamma(\theta) = \int_{-\infty}^{+\infty} f_C(u) S_T(u) du - \tau = \int_0^{+\infty} \frac{1}{\theta \nu u^{1/\nu}} \Gamma\left(\frac{1}{\nu}, u\theta^\nu \right) - \tau.$$

Enfin si $\nu = 1, T \sim \mathscr{E}(\lambda_x)$, la formule de $\gamma(\theta)$ reste la même. Mais pour un *sparse rate sr* à 0, $T \sim \mathscr{E}(\alpha)$, celle-ci se simplifie en :

$$\begin{aligned}\gamma(\theta) = \int_0^{+\infty} \mathbb{P}(\delta = 1|u, \nu, \theta) f_{\lambda_i}(u) du - \tau &= \int_0^{\theta} \frac{1}{\theta} exp(-\alpha u) du - \tau \\ &= \frac{1}{\theta}\left[-\frac{exp(-\alpha u)}{\alpha}\right]_0^{\theta} - \tau \\ &= (1 - exp(-\alpha\theta)) - \tau.\end{aligned}$$

Schémas de simulation. Pour chaque scénario, nous simulerons $M = 100$ jeux de données avec un taux de censure de $\tau = 0.5$. Nous choisirons $a = 0$ et $m_0 = 10$. Nous ferons varier :

— le gain de médiane pour l'augmentation de une unité : $m_1 - m_0 \in \{0.5, 1\}$,
— la taille de l'échantillon : $n \in \{100, 150, 200, 500\}$,
— le taux de variables actives : $sr \in \{0, 0.25, 0.5\}$.

Résultats sur l'optimisation des hyperparamètres et la validation croisée

Le graphique A de la figure 4.6 correspond à un taux de variables actives, le *sparse rate*, de 0, c'est-à-dire qu'il n'y a pas de covariable informative, pour 4 tailles d'échantillon différentes puisque N varie de 100 à 500 (en lignes). Pour chacun de ces 4 scénarios ont été simulés 100 jeux de données, analysés par Cox Boost en validation croisée en 2 et 5 blocs (en colonne), sans optimisation des hyperparamètres en vert, Halving en rouge, et Optuna en bleu.

Comme espéré, le C de Harrell par validation croisée est bien centré sur 0.5, donc peu biaisé (nous avons utilisé la *nested cross-validation*). Comme attendu, quand N augmente la variabilité diminue. Sur le graphique B, le taux de variables actives est passé de 0 à 0.25, et maintenant qu'il y a des covariables actives, il a fallu fixer la taille de l'effet des covariables, ici avec *Gainmed* à 0.5 ($m_1 - m_0$) on passe d'une médiane de survie de 10 à une médiane de 10.5 quand on augmente d'une unité une covariable. Le graphique montre un C de Harrell qui augmente, et une légère efficacité de l'optimisation des hyperparamètres pour $N = 500$. Sur le graphique C, le taux de variables actives passe de 0.25 à 0.5. Les commentaires précédents restent vrais et plus nets avec un C de Harrell plus élevé. En comparant les graphiques B et D avec le taux de variables actives à 0.25 et le gain de médiane qui passe de 0.5 à 1, comme attendu, un effet des covariables plus marqué est visible, le C de Harrell augmente un peu, il était par exemple pour N=500 en dessous de 0.6 en B et au dessus en D. Pour un taux de variables actives à 0.5 (graphiques C et D), les tendances sont plus marquées, le C de Harrell augmente avec N, mais n'augmente que très légèrement avec les validations croisées de 2 à 5 blocs, et Halving semble légèrement meilleur que Optuna.

Les résultats en termes de temps de calcul (non montré) sont que la validation croisée 5 blocs n'améliore que très légèrement mais est 2 à 3 fois plus longue (entre 15 secondes et 3h selon les jeux de données). L'efficacité de l'optimisation des hyperparamètres est indéniable, Halving faisant légèrement mieux que Optuna mais beaucoup plus gourmande en temps de calcul : Optuna met en moyenne 20 fois plus de temps que sans optimisation et Halving 170 fois plus de temps que sans optimisation.

Sur la figure 4.7, les AUC_{Imp} sont très proches de 0.5, ce qui montre les lacunes des mesures d'importance des variables à détecter les variables simulées actives, et ce même quand la performance du modèle en termes de classement est correcte. On peut voir en haut à droite un C de Harrell à 0.8 et un AUC_{Imp} à 0.51. Une tendance pourrait peut-être se deviner pour N=500, mais encore une fois les AUC_{Imp} restant si bas, ces simulations ne permettent pas de montrer quoi que ce soit. Cependant, un choix méthodologique

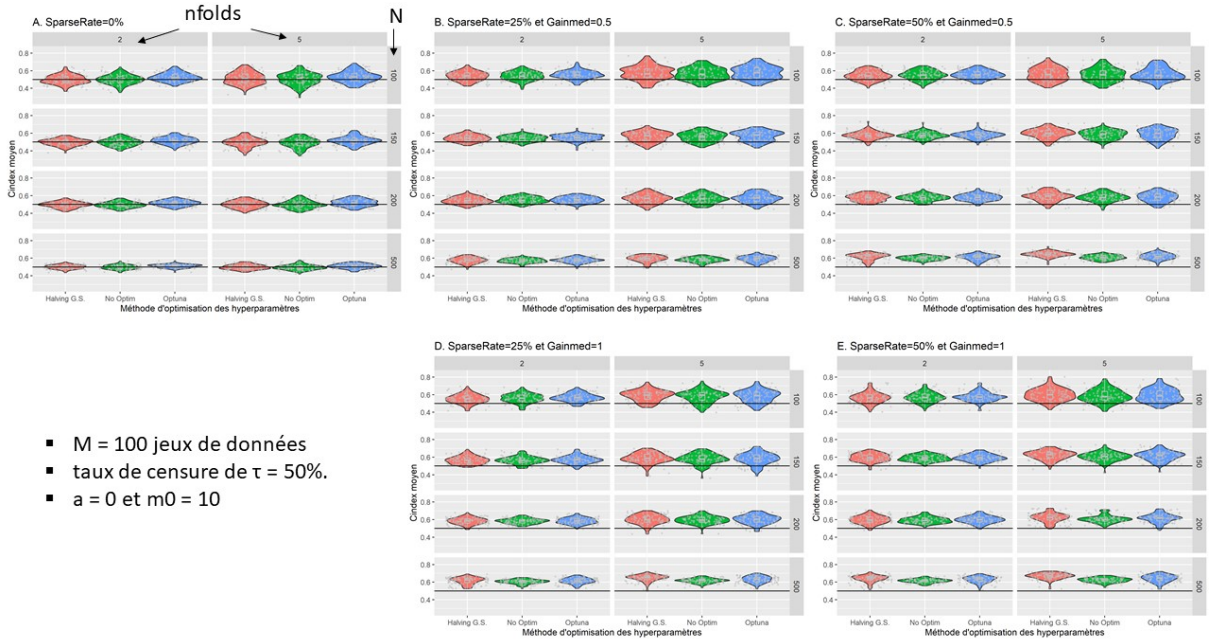


FIGURE 4.6 – Distribution des C_{Index} moyens en fonction des AUC_{Imp} sur les 2 (respectivement 5) blocs de la validation croisée pour les $M = 100$ jeux de données simulées, $n \in \{100, 150, 200, 500\}$, $m_1 - m_0 \in \{0.5, 1\}$, $\text{sr} \in \{0.25, 0.5\}$.

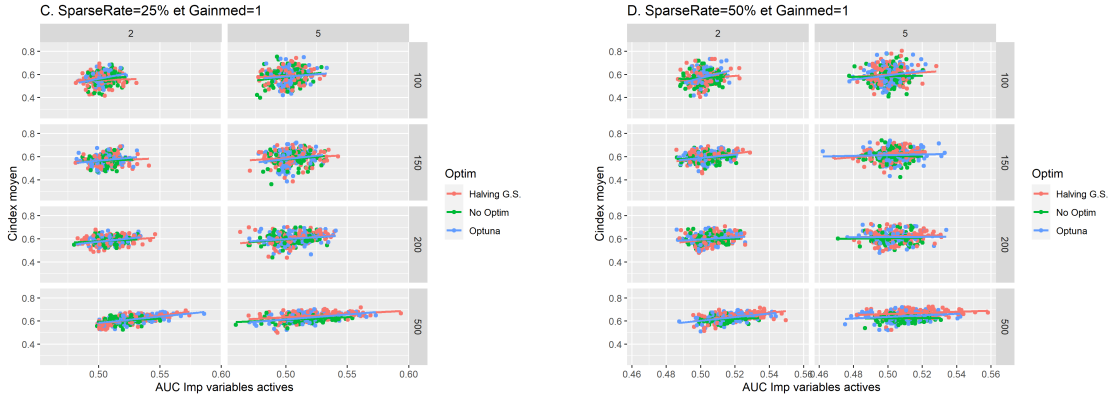


FIGURE 4.7 – Distribution des C_{Index} moyens sur les 2 (respectivement 5) blocs de la validation croisée pour les $M = 100$ jeux de données simulées, $n \in \{100, 150, 200, 500\}$. En C. $m_1 - m_0 = 0.5 et \text{sr} = 0.25$. En D. $m_1 - m_0 = 1 et \text{sr} = 0.5$.

a amplifié le problème, la méthode Cox Boost choisie était celle qui permet de mettre des coefficients à 0, le nombre de coefficients effectivement mis à 0 étant très élevé, énormément d'Importance de variables sont également exactement à 0, ce qui a une conséquence sur le calcul de l'AUC_{Imp}. Ces résultats nous motivent à continuer les investigations et à continuer l'étude en comparant avec la méthode Cox Boost basée sur les arbres.

4.2.4 Application sur un jeu de données réelles

Le jeu de données **NSBCD** comprend les profils d'expression de 549 gènes "intrinsèques" de 115 tumeurs malignes du sein. Ce jeu de données pouvait être téléchargé sur http ://user.it.uu.se/ liuya610/ dans les années qui ont suivi la publication de l'article [127] et sont maintenant disponibles sur : https ://user.it.uu.se/ kripe367/survlab/download.html.

La figure 4.8 montre que l'optimisation des hyperparamètres par Halving Grid Search donne de très bonnes performances (C de Harrell proches de 0.75 en validation croisée à 5 blocs) avec les 3 méthodes, mais CoxBoost fait exploser le temps de calcul. Les hyperparamètres standard de la méthode Survival SVM semblent adaptés à ce jeu de données qui ne nécessite donc pas réellement d'optimisation des hyperparamètres (les deux croix rouges et noires sont très proches en haut à gauche du graphique). Optuna n'est par ailleurs pas du tout performant ici (croix vertes avec des C_{Index} très faibles). CoxBoost est le plus long en temps de calcul en cas d'optimisation des hyperparamètres.

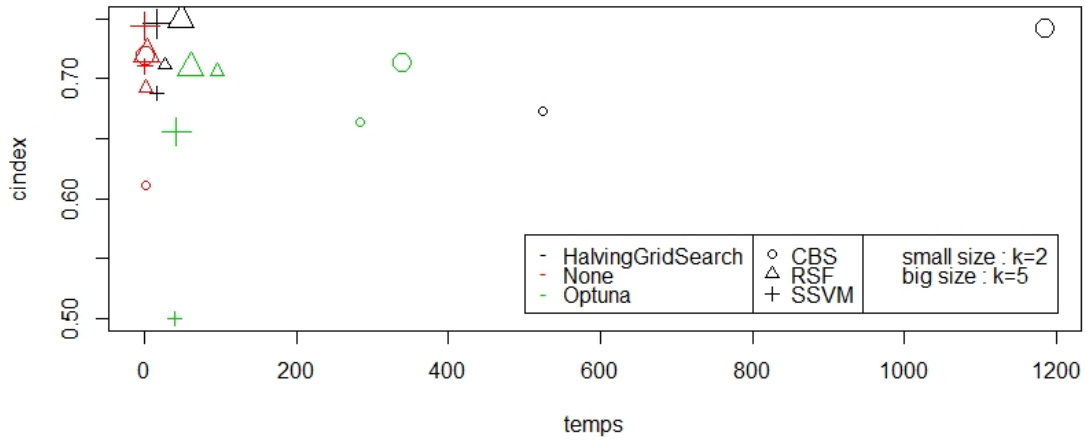


FIGURE 4.8 – Performance en termes de C de Harrell des méthodes CoxBoost, Random Survival Forest, et Survival SVM par validation croisée 2 et 5 blocs et optimisation ou non des hyperparamètres sur le jeu de données réelles NSBCD.

4.3 Événements récurrents en grande dimension

Nous nous sommes intéressés jusqu'ici à des données qualifiées de données de survie concernant le temps d'apparition d'un événement (time-to-event). Dans la même lignée, il existe un autre type de données très classique dans le domaine médical, ce sont les événements répétés au fil du temps, comme des hospitalisations ou des rechutes de cancer. Dans les essais cliniques ou dans le monde réel, l'analyse de survie se concentre habituellement sur la modélisation du temps jusqu'à la première occurrence de l'événement. Néanmoins, les variables peuvent avoir un impact variable sur le premier événement et sur les événements qui suivent, et la spécificité de censure persiste (figure 4.9).

Deux principaux défis se posent lors de l'analyse des événements récurrents. Tout d'abord, l'hétérogénéité interindividuelle apparaît car certains sujets peuvent être plus susceptibles que d'autres de vivre

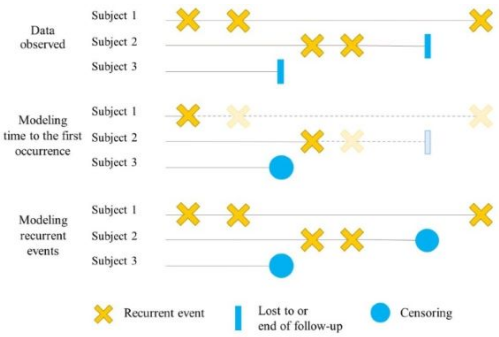


FIGURE 4.9 – Figure 1 du 1er article de thèse de Juliette Murris soumis à Biostatistics and Epidemiology [16]. Illustration de la spécificité des données de type évènements récurrents.

l'événement. Deuxièmement, les événements pour un individu ne sont pas indépendants, ce qui augmente l'hétérogénéité intra-individuelle. Ces questions ont été traitées statistiquement au moyen de deux approches, les modèles marginaux et conditionnels. Les modèles marginaux comportent une moyenne implicite des événements récurrents antérieurs. Les modèles conditionnels peuvent dépendre de l'histoire des événements.

Bien évidemment, tout comme dans la partie précédente, les difficultés liées à la grande dimension arrivent à nouveau. Et la thèse de Juliette Murris que je co-encadre se situe dans ce contexte.

4.3.1 Revue de la littérature

Un recensement systématique des articles a été effectué sur PubMed pour des études potentielles publiées dans des revues jusqu'en novembre 2021, selon les recommandations de Cochrane [148, 149]. Des recherches manuelles ont ensuite été effectuées au moyen de moteurs de recherche et de conférences pertinents [150]. Les critères d'inclusion étaient des études méthodologiques ou observationnelles qui analysaient tout résultat récurrent. Des méthodes ont été envisagées si la publication mentionnait que des données dans un cadre à haute dimension étaient utilisées ou si des techniques d'apprentissage automatique étaient utilisées. Les critères d'exclusion étaient toute approche bayésienne et tout plan d'essai clinique. La requête Pubmed comprenait (sans s'y limiter) les termes clés suivants (et les termes MeSH connexes) : « récidive », « analyse de survie », « grande dimension » et « apprentissage automatique ». Deux évaluateurs ont évalué l'admissibilité des publications de façon indépendante et ont discuté de toute divergence. Un suivi des citations faites dans l'article x ou réciproquement dans les articles où l'article x a été cité, a été effectué pour éviter de manquer toute documentation pertinente. Les données ont été résumées de façon descriptive en fonction des catégories suivantes : caractéristiques de publication et d'étude, approches statistiques/d'apprentissage automatique utilisées et application des données.

L'extraction a mené à l'identification de 176 résultats de recherche électronique dans la base de données Pudmed (figure 4.10). Dans l'ensemble, après avoir confirmé le résultat d'intérêt relatif à la récurrence, la principale raison d'exclusion était la non-considération des événements récurrents en tant que moment d'événement pour chaque événement. La récurrence a été considérée comme un classificateur (19/176), comme un résultat de survie sans récurrence (23/176), ou comme un événement de la première fois (29/176). C'est peut-être l'illustration de la prudence des auteurs lorsqu'ils traitent d'événements récurrents de grande envergure, car il n'existe pas de lignes directrices ou de recommandations publiées à ce jour. De plus, trois articles en texte intégral n'ont pu être examinés, car ils n'étaient pas disponibles.

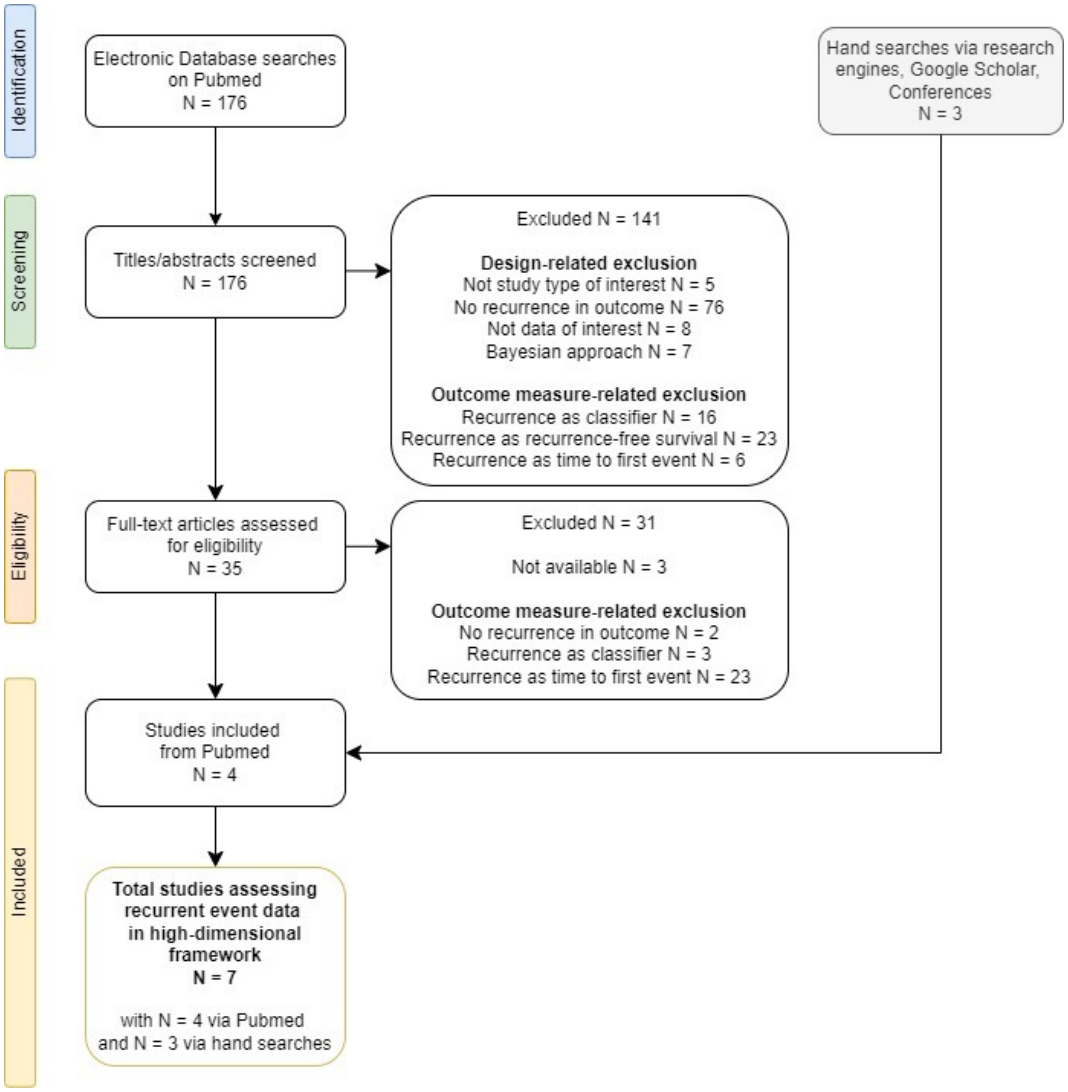


FIGURE 4.10 – Figure 2 du 1er article de thèse de Juliette Murris soumis à Biostatistics and Epidemiology [16]. Flowchart des publications incluses via Pubmed.

Après un examen approfondi du titre, du résumé et du texte intégral, quatre publications ont été incluses à partir de la recherche dans la base de données électronique. Trois autres documents issus d'une recherche manuelle ont été identifiés.

Au total, sept publications pertinentes ont été sélectionnées :
- quatre études méthodologiques,
- un document d'application et
- deux revues de la littérature.

Deux articles décrivant les algorithmes d'apprentissage pour les stratégies de sélection des variables ont été identifiés. Tout d'abord, Wu (2012) a mis l'accent sur l'accélération de l'estimation du coefficient avec un algorithme de descente coordonné et la pénalisation de la vraisemblance partielle [151]. Zhao

(2018) a fourni une prolongation de la pénalisation des crêtes pour l'estimation et la sélection simultanées des variables [152]. Gupta (2019) et Jing (2019) ont développé des extensions de réseaux neuronaux profonds pour l'analyse de la récurrence, respectivement [153, 154]. En outre, Kim (2021) a proposé un document d'application visant à estimer le temps entre deux récidives de cancer du sein [155]. Cependant, la méthodologie utilisée dans ce dernier était une extension d'un réseau neuronal récurrent qui n'a malheureusement pas été publié dans une revue à comité de lecture. Enfin, deux revues de la littérature ont mentionné des algorithmes d'apprentissage pour l'analyse des données sur les événements récurrents, même si aucune d'entre elles ne fournissait de lignes directrices [156, 157].

4.3.2 Modélisation des évènements récurrents

On note $i = 1, ..., n$, avec n le nombre de sujets et $\mathbf{X} \in \mathbb{R}^{n \times p}$ la matrice des covariables pour tous les sujets. Soit $\mathbf{X}_i$ un vecteur de covariables de dimension p, β les coefficients de régression associés, $\lambda_0(t)$ la fonction de risque de base, $Y_i(t)$ un indicateur permettant de savoir si le sujet i est à risque au temps t, $\delta_i = 1$ lorsque le sujet a subi au moins un événement (sinon 0). Soit C_i le temps de suivi ou de censure. $N_i^*(t)$ désigne le nombre d'événements sur l'intervalle $[0, t]$. Dans le paragraphe "L'aire sous la courbe ROC temps-dépendante" au début du chapitre, nous avions signalé que le temps de survie peut être considéré comme un résultat binaire variable dans le temps en se concentrant sur la représentation du processus de comptage $N_i^*(t) = \mathbf{1}_{\{T_i < t\}}$. Nous avons choisi de garder la même notation $N_i^*(t)$ ici, même si $N_i^*(t)$ ne se limite plus aux valeurs 0 ou 1.

Modèles statistiques standards

Les modèles d'Andersen-Gill (AG), de Prentice, William et Peterson (PWP), de Wei-Lin-Weissfeld (WLW) et de fragilité ont été développés comme des extensions du modèle de Cox [158, 159, 160, 161, 162]. Ces méthodologies sont des modèles couramment utilisés pour traiter les données d'événements récurrents. Leurs caractéristiques sont résumées dans le tableau 4.2.

La modélisation d'événements récurrents lors de l'utilisation d'approches statistiques standard nécessite de définir la notion d'individus à risque et d'échelle de temps.

Ensemble d'individus à risque Les modèles statistiques standard décrits ne considèrent pas les individus à risque de la même manière. Cela induit une gestion préalable des données pour une application appropriée.

- L'ensemble des individus à risque pour l'événement k comprenant les individus qui étaient à risque pour l'événement. Il existe différentes définitions de l'ensemble des individus à risque, principalement basées sur la fonction de risque de base ;
- L'ensemble non restreint, dans lequel chaque sujet peut être à risque pour n'importe quel événement, quel que soit le nombre d'événements présentés, à intervalles réguliers ;
- L'ensemble restreint ne contenant que les intervalles de temps du $k^{\text{ème}}$ événement des sujets qui avaient déjà présenté $k - 1$ événements. ;
- L'ensemble semi-restreint contenant pour le $k^{\text{ème}}$ événement les sujets qui ont eu $k - 1$ ou moins d'événements.

Modèle	Composants et spécificités
AG	Modèle conditionnel, tient compte du processus de comptage en tant qu'échelle de temps et ensemble non restreint pour les sujets à risque Les événements récurrents pour chaque individu sont indépendants et partagent une fonction de risque de base commune Fonction de risque : $\lambda_i(t) = Y_i(t)\lambda_0(t)\exp(\beta^t X_i)$
PWP	Modèle conditionnel, processus de comptage comme échelle de temps et ensemble restreint pour les sujets à risque AG stratifié, la strate k rassemble tous les $k$èmes événements des individus $\lambda_{ik}(t) = Y_i(t)\lambda_{0k}(t)\exp(\beta_k^t X_i)$
WLW	Modèle marginal, également stratifié, échelle de temps calendaire et ensemble semi-restreint pour les sujets à risque Dépendance intra-sujet $\lambda_{ik}(t) = Y_i(t)\lambda_{0k}(t)\exp(\beta_k^t X_i)$
Frailty	Extension du modèle AG Terme aléatoire z_i pour chaque individu afin de tenir compte des caractéristiques non observables ou non mesurées $\lambda_i(t) = Y_i(t)\lambda_0(t)z_i\exp(\beta^t X_i)$

AG = Andersen-Gill ; PWP = Prentice, William et Peterson ; WLW = Wei-Lin-Weissfeld.

TABLE 4.2 – Modèles statistiques standards pour l'analyses des événements récurrents

Échelles de temps Les délais sont également des éléments clés à prendre en compte lors de la gestion des données. Les trois délais les plus courants sont les suivants :

- Le temps calendaire, dans lequel les temps indiquent le temps écoulé depuis la randomisation/le début de l'étude jusqu'à ce qu'un événement se produise ;
- Le temps de passage, ou échelle d'attente, remet le temps à zéro lorsqu'un événement se produit, c'est-à-dire qu'il correspond au temps écoulé depuis le dernier événement précédemment observé ;
- Le processus de comptage est construit selon le temps calendaire, bien qu'il permette des inclusions tardives et/ou des censures.

La figure 4.11 illustre ces échelles de temps.

Algorithmes d'apprentissage pour la sélection de variables

Une approche courante pour relever le défi de la haute dimension (et le traitement de la multicolinéarité des données) est la sélection de variables par les modèles de pénalisation Lasso et Ridge qui ont été explicités dans le chapitre 2. Les deux approches de pénalisation ont été étendues aux modèles de Cox dans le cadre standard de l'analyse de survie [163, 164], ce que nous avons par ailleurs déjà évoqué lors de l'explication de la méthode Cox Boost.

Une extension de ces méthodes aux événements récurrents a été proposée dans l'article [152] en développant la régression *Broken Adaptive Ridge* (BAR). La première itération consiste en un modèle L_2 pénalisé.

$$\hat{\beta}^{(0)} = \text{argmin}_\beta(-2\ell_{mod}(\beta) + \xi_n \sum_{j=1}^{p} \beta_j^2), \xi_n \geq 0,$$

où $\ell_{mod}(\beta)$ est la vraisemblance modifiée pour les modèles d'événements récurrents et ξ_n est le paramètre de pénalisation.

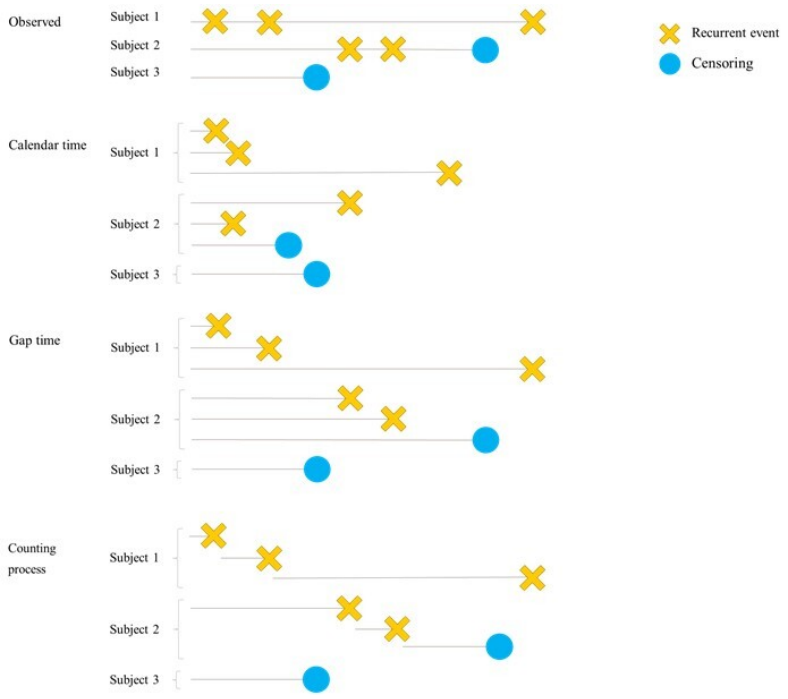


FIGURE 4.11 – Figure 7 du 1er article de thèse de Juliette Murris soumis à Biostatistics and Epidemiology [16]. Échelles de temps dans les analyses d'événements récurrents.

Si $\xi_n \geq 0$, alors c'est une pénalité Ridge, et si $\xi_n = 0$ alors $\beta^{(0)}$ n'est pas pénalisé. Nous mettons à jour à chaque itération ω :

$$\hat{\beta}^{(\omega)} = \text{argmin}_\beta(-2\ell_{mod}(\beta) + \theta_n \sum_{j-1}^{p} \frac{\beta_j^2}{\hat{\beta}_j^{(\omega-1)}}), \omega \geq 1,$$

où θ_n est le deuxième hyperparamètre de pénalisation.

Les estimations BAR sont définies par $\hat{\beta} = \lim_{k\to\infty} \hat{\beta}^{(\omega)}$. La validation croisée est recommandée pour optimiser les valeurs des hyperparamètres ξ_n et θ_n. Selon l'article [165], les estimations ne sont pas sensibles aux variations de ξ_n et l'optimisation peut n'être effectuée que sur θ_n. En l'absence d'une mesure unique consensuelle sur la validation croisée pour la modélisation des événements récurrents, deux valeurs de θ_n ont été étudiées dans ce travail, considérant ainsi deux modèles distincts.

Du C de Harrell au C index de Kim

Kim a proposé une mesure de concordance entre les nombres d'événements observés et prédits sur un intervalle de temps d'observations [166]. Il s'agit de la proportion de paires d'individus pour lesquelles la prédiction du risque et le nombre d'événements observés sont concordants :

$$\hat{\mathscr{C}}_{rec} = \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} \mathbf{1}_{N_{i'}^*(C_i \wedge C_{i'}) > N_{i'}^*(C_i \wedge C_{i'})} \mathbf{1}_{({}^t\beta X_i > {}^t\beta X_{i'})}}{\sum_{i=1}^{n}\sum_{i'=1}^{n} \mathbf{1}_{N_i^*(C_i \wedge C_{i'}) > N_{i'}^*(C_i \wedge C_{i'})}},$$

où $N_{i'}^*(C_i \wedge C_{i'})$ est le nombre d'événements pour l'individu i' entre le temps 0 et le minimum entre la

durée d'observation de l'individu i et l'individu i'.

Cette extension de l'indice C implique :
- Deux individus sont comparables jusqu'à la durée minimale du suivi ;
- Une paire contribue au dénominateur si les deux nombres d'événements ne sont pas égaux.

Tout comme pour le C de Harrell, un score proche de 1 indique une meilleure performance du modèle.

4.3.3 Simulations

Génération de données

Les hypothèses suivantes ont été formulées :
- Les variables actives sont continues et ont toutes le même effet (non nul) ;
- Les variables ne varient pas dans le temps ;
- Les individus étaient à risque de façon continue jusqu'à la fin du suivi ;
- La censure n'est pas informative.

La génération de la matrice des covariables, $X \sim \mathcal{N}_m(\mu, \sigma(\rho))$. $\mu = (\mu_1 ... \mu_p) = (a...a)$ et $\sigma(\rho))$ est la matrice de covariance avec une structure de corrélation autorégressive et $\rho \in (0, 1)$. Les coefficients $\beta = (\beta_1 ... \beta_p) = (b, ..., b, 0, ..., 0)$ étaient associés aux covariables p. Les coefficients m étaient égaux à une constante $b \in \mathbb{R}$ (la valeur des coefficients actifs) et les coefficients $p - m$ étaient égaux à zéro. Le taux de variables actives était décrit par $\frac{m}{p}$. La fonction de risque de base suivait une distribution de Weibull avec un paramètre d'échelle $\alpha > 0$ et un paramètre de forme $\gamma > 0$, et $\lambda_0(t) = \alpha\gamma t^{(\gamma-1)}$. La fonction de risque de base cumulée peut être exprimée par $\Lambda_0(t) = \int_0^t \lambda_0(s)ds = \alpha t^\gamma$. Par conséquent, la fonction de risque cumulé peut être exprimée par $\Lambda(t|X_i) = \Lambda_0(t)\exp({}^t\beta X_i)$. La fonction de hasard de base conditionnelle a ensuite été définie comme $\tilde{\Lambda}_t(u) = \tilde{\Lambda}^i(u|T_{i-1} = t) = \Lambda(u + t) - \Lambda(t)$. Un terme de fragilité z_i i.i.d. a été incorporé pour tenir compte de l'hétérogénéité.

Pour maintenir les taux de censure, les individus censurés ont été tirés au sort (la censure n'est pas informative), conformément à l'article [167].

L'algorithme de [168] a été appliqué pour simuler les temps d'événements k pour chaque sujet i :

$$t_{i,1} = \Lambda^{-1}(t)(-\ln(\epsilon_1)),$$
$$t_{i,k+1} = t_{i,1} + \tilde{\Lambda}^{-1}_{i,t_k}(-\ln(\epsilon_{k+1})),$$

avec $\epsilon_k \sim U([0, 1])$.

Une répartition Train-Test a été utilisée avec une distribution de 70-30%. Les jeux de données ont été générés avec :
- N = 100 sujets
- taux de censure de 20%
- $\rho = 0.7$
- $b = 0.15$
- $\alpha = 1$ et $\gamma = 2$.
- $z \sim Gamma(0.25)$

Schémas de simulation

Les scénarios comprennent des variations du nombre de covariables p = 25, 50, 100, 150 et 200 et du taux de variables actives *sparse rate* = 0%, 25% et 50%. Pour chacun des 15 scénarios, 100 jeux de données ont été générés pour tenir compte de la variabilité.

Résultats

Les indices C moyens ont été calculés sur l'ensemble des 15 scénarios (figure 4.12). Comme prévu, les modèles standard échouent dès que $p > n$. Alors que l'on s'attendait également à ce que les indices C se situent autour de 0.5 lorsque le taux de variables actives était nul, ils augmentent à mesure que le taux de variables actives augmente.

La meilleure performance a été obtenue en utilisant le modèle de fragilité. Les autres modèles ont montré des tendances similaires, à l'exception des modèles WLW. Les indices C de ce modèle sont restés autour de la valeur de 0.5 (et même en dessous) quel que soit le scénario.

L'indice C de Kim était plus stable pour les différents nombres de covariables et taux de variables actives, même s'il avait tendance à diminuer lorsque le nombre de covariables augmentait avec un taux de variables actives de 50%. Une petite différence entre les valeurs de pénalité a été remarquée, les modèles pénalisés à 0.05 et à 0.1 suivant des tendances similaires.

Cette étude est la première à comparer des méthodes standard, des algorithmes de sélection de variables et un réseau de neurones profond pour modéliser des événements récurrents dans un cadre de grande dimension.

L'utilisation de technologies d'intelligence artificielle (IA) en santé est en plein essor. Ces systèmes sont généralement conçus pour prévenir l'apparition d'événements à l'hôpital, dans une maison de retraite ou en ambulatoire, etc. Si ces événements sont susceptibles de se produire de manière répétée et une grande quantité de données éventuellement informatives est disponible, une analyse approfondie, robuste et appropriée des événements récurrents en grande dimension est alors cruciale [169].

Dans l'ensemble, ce travail soulève de nombreuses préoccupations concernant l'analyse des données d'événements récurrents dans des contextes de grande dimension et souligne la nécessité actuelle de développer de nouvelles approches et d'évaluer leurs performances de manière pertinente.

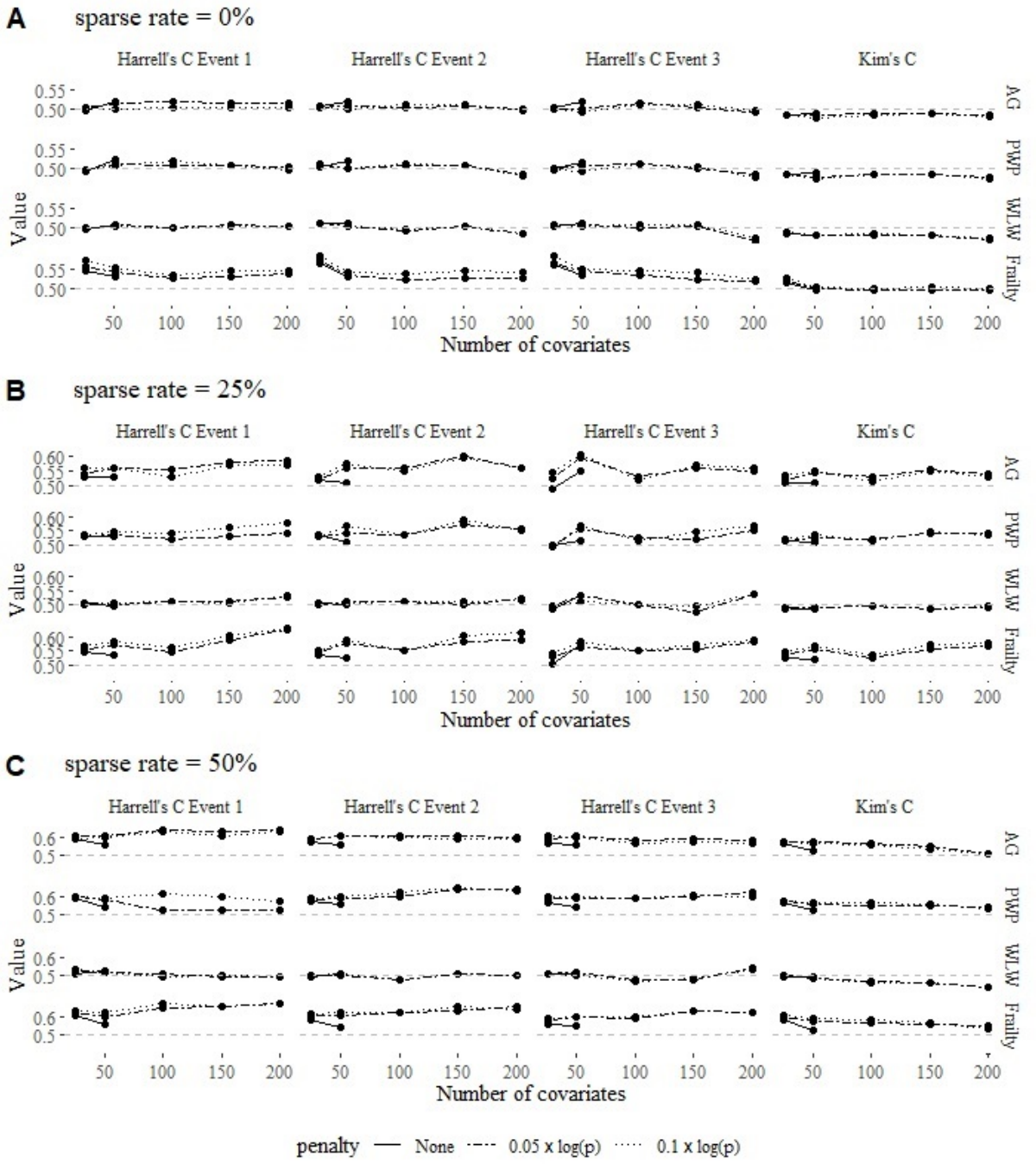


FIGURE 4.12 – Figure 4 du 1er article de thèse de Juliette Murris soumis à Biostatistics and Epidemiology [16]. Impact du numbre de covariables sur les C-index moyens avec un taux de variables actives égal à 0% (A), 25% (B) et 50% (C).

Note de lecture : *p le nombre de covariables. Pour chaque taux de variables actives, modèle et pénalité, les indices C moyens des 100 jeux de données simulés ont été représentés en fonction du nombre de covariables. Les pénalités étaient égales à 0 (non pénalisé),* $0.05 \times log(p)$ *et* $0.1 \times log(p)$*, respectivement. Les modèles statistiques standards non pénalisés ne convergeaient pas dès que $p > n$, les performances n'étaient donc pas disponibles. AG : Andersen-Gill ; PWP : Prentice, William, et Peterson ; WLW : Wei-Lin-Weissfeld.*

4.4 Perspectives

4.4.1 Comparaison de méthodes d'analyse de survie en grande dimension

Les résultats présentés en partie 4.2.3 et publiés dans les actes des journées de statistique de la SFDS 2022 [14] ont vocation à être étendus dans un travail de collaboration toujours avec Magalie Fromont, Valérie Garès, Juliette Murris et Sandrine Katsahian.

D'autres critères d'évaluation

Du C moyen au C sur échantillon reconstruit. Lorsque l'objectif n'est pas seulement le classement mais la prédiction elle-même, un critère de performance de type MSE pour une variable réponse quantitative ou taux de mal classés pour une variable réponse qualitative binaire sera préféré au C de Harrell. Il s'agit d'un critère de risque, classiquement utilisé en modèles de régression ou de discrimination (sur des données non censurées).
Considérons une partition $\{\mathscr{I}_1, \dots \mathscr{I}_K\}$ de $\{1, \dots, n\}$. Pour chaque $k = 1, \dots, K$, la règle de prédiction g est construite avec l'échantillon d'apprentissage $\{(x_i, y_i), i \in \{1, \dots, n\} \setminus \mathscr{I}_k\}$, on la note $\widehat{g}^{n,k}$.
Le critère $\mathrm{MSE}_{n,k}(\widehat{g}^{n,k})$ est calculé sur l'échantillon test $\{(x_i, y_i), i \in \mathscr{I}_k\}$.
Considérons ensuite le risque moyen sur les K blocs :

$$\mathrm{MSE}_m = \frac{1}{K}\sum_{k=1}^{K} \mathrm{MSE}_{n,k}(\widehat{g}^{n,k}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}(y_{ik} - \hat{y}_{ik})^2.$$

Prenons maintenant l'échantillon reconstruit des n individus avec leur prédiction obtenue par le modèle qui a été construit sans eux (et ceux de leur bloc) $\{(x_i, \hat{y}_i), i \in \mathscr{I}_k\}$ avec $\hat{y}_i$ obtenu par $\widehat{g}^{n,k}$, en mettant bout à bout ses K ensembles d'individus, et calculons :

$$\mathrm{MSE}_r = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Le seul opérateur étant la somme, ces deux approches donnent le même résultat, même chose pour les proportions de bien classés où la moyenne des proportions au sein des blocs MSE_m est égale à la proportion sur échantillon reconstruit MSE_r. Pour le C de Harrell c'est très différent, puisqu'il est évident que la moyenne des comparaisons deux à deux au sein des blocs n'est pas égale à la moyenne des comparaisons deux à deux sur l'échantillon reconstruit. De plus, nous avions précédemment fait remarquer que plus le nombre de blocs augmente, plus la taille de l'échantillon d'apprentissage est grande, plus celle de l'échantillon test est petite, et dans le cas du "leave-one-out" avec un seul individu par bloc, le C de Harrell ne peut pas être calculé. Seul le C de Harrell sur échantillon reconstruit peut alors être calculé. Quand n est très grand le leave-one-out (ou même un nombre de blocs K assez grand) n'est pas une option à cause du temps de calcul qui explose, dans notre cas où n est petit et p très grand, il semble très intéressant d'étudier de près ces deux approches pour tenter de dessiner des recommandations pour l'utilisation du critère le moins biaisé.
Une autre approche plus chronophage est le leave-two-out qui construit autant d'échantillons d'apprentissage que de combinaisons de 2 individus à sortir de ces échantillons d'apprentissage pour former les échantillons test, soit $n(n-1)$. On peut alors calculer la moyenne de ces $n(n-1)$ C de Harrell. Par contre, n'étant dans ce cas plus sur une validation croisée à K blocs, envisager un échantillon réconstruit n'aurait pas de sens puisque plusieurs échantillons test sont formés des mêmes individus.

C de Uno. L'estimation du C-index est basée sur les paires (i, j) utilisables mais ne tient pas compte de la probabilité d'être censuré, ce qui peut être corrigé [170].

Rappelons la formule du C de Harrel :

$$\mathrm{C}_{\mathrm{Index}} = \frac{\sum_{i=1}^{n} \sum_{i'=1, i \neq i'}^{n} \mathbf{1}_{\{T_i^O > T_{i'}^O\}} . \mathbf{1}_{\{M_{i'} > M_i\}} . \delta_{i'}}{\sum_{i=1}^{n} \sum_{i'=1, i \neq i'}^{n} \mathbf{1}_{\{T_i^O > T_{i'}^O\}} \delta_{i'}}.$$

Uno modifie cet indice de concordance en pondérant les indicatrices de paires concordantes et de paires utilisables. Soit $G(.)$ est la fonction de survie de la censure. Les poids sont des estimations de Kaplan Meier de la distribution de survie de la censure (c'est-à-dire la distribution de probabilité pour les individus de ne pas être censurés avant le temps t), notés $\hat{G}(t)$.

$$\mathrm{C}_{\mathrm{Uno}} = \frac{\sum_{i=1}^{n} \sum_{i'=1, i \neq i'}^{n} \left\{G(T_{i'}^O)\right\}^{-2} \mathbf{1}_{\{T_i^O > T_{i'}^O\}} . \mathbf{1}_{\{M_{i'} > M_i\}} . \delta_{i'}}{\sum_{i=1}^{n} \sum_{i'=1, i \neq i'}^{n} \left\{G(T_{i'}^O)\right\}^{-2} \mathbf{1}_{\{T_i^O > T_{i'}^O\}} \delta_{i'}}$$

Nous avons calculé le C de Uno pour la comparaison des 3 méthodes sur le jeu de données réelles NSBCD, la figure 4.13 montre que l'optimisation des hyperparamètres par Halving Grid Search donne un C de Uno supérieur à 0.75 en validation croisée à 5 blocs avec les 3 méthodes, et se distingue maintenant de la non optimisation des hyperparamètres, qui est cependant globalement meilleure que Optuna, ce qui s'explique sûrement par le fait que les hyperparamètres sont optimisés sur le C de Harrell pour construire le modèle, même quand nous étudions les trois critères de performance (C de Harrell, C de Uno et IBS). CoxBoost fait toujours exploser le temps de calcul en cas d'optimisation des hyperparamètres.

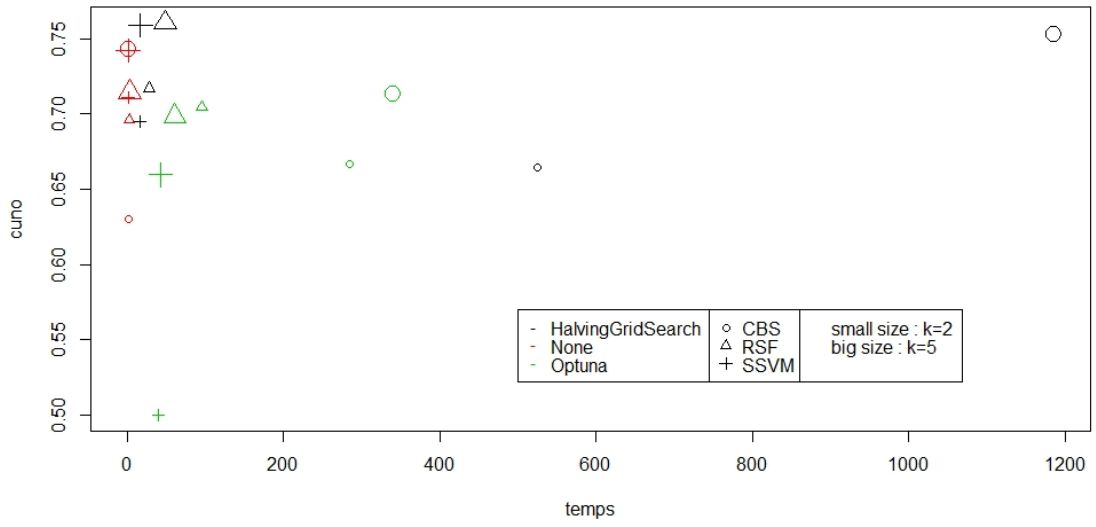


FIGURE 4.13 – Performance en termes de C de Uno des méthodes CoxBoost, Random Survival Forest, et Survival SVM par validation croisée 2 et 5 blocs et optimisation ou non des hyperparamètres sur le jeu de données réelles NSBCD.

Score de Brier intégré (IBS). Le pouvoir de classement, ou de concordance, et le pouvoir prédictif sont tous deux des éléments essentiels pour évaluer la performance globale d'un modèle. Les indices C

décrits ci-dessus (Harrell et Uno) sont des critères de performance de classement. Le score de Brier (BS) est un critère de performance prédictive.
Dans un cadre de discrimination binaire classique sans censure, ce score est donné par

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$$

où p_i est la probabilité prédite par le modèle d'avoir l'événement pour l'individu i, et $o_i = 1$ si l'individu i a eu l'événement,$o_i = 0$ sinon. Un modèle avec un score BS plus faible donnera de meilleures prédictions et sera considéré comme bien calibré puisque le BS est proche de 0.

Nous pouvons adapter ce score au cadre de survie - d'abord sans censure. Dans ce cas, $BS(t)$ compare au temps t la prédiction de survie basée sur le modèle avec le statut de survie observé [171] :

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}(0 - \hat{S}(t|x_i))^2 & \text{si } t_i^O \leq t \\ (1 - \hat{S}(t|x_i))^2 & \text{si } t_i^O > t\end{cases}$$

Avec $\hat{S}$ l'estimateur de la fonction de survie. Une faible probabilité de survie sera prédite pour un individu ayant subi un événement avant le temps t, alors qu'une forte probabilité de survie est attendue lorsque l'individu subit l'événement après t.
Voyons maintenant ce qui se passe avec la censure des données de survie. Graf et al. introduisent la pondération dans le score de Brier en utilisant des pondérations de censure à probabilité inverse (IPCW), évitant ainsi de biaiser la moyenne de la population [171, 172]. Comme précédemment pour le C de Uno, les poids sont des estimations de Kaplan Meier de la distribution de survie de la censure notés $\hat{G}(t)$. La BS s'écrit alors

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{\hat{G}(t_i)}(0 - \hat{S}(t|z_i))^2 & \text{if } t_i \leq t, \delta_i = 1 \\ \frac{1}{\hat{G}(t)}(1 - \hat{S}(t|z_i))^2 & \text{if } t_i > t\end{cases}$$

Cela signifie que les individus subissant l'événement avant le temps t ont un poids de $\hat{G}(t_i)$ alors que les individus subissant l'événement ou la censure après le temps t ont un poids de $\hat{G}(t)$. Comme l'estimation de la distribution de survie de la censure, $\hat{G}(t)$ est une fonction décroissante. Pour cette raison, on s'attend à ce que $1/\hat{G}(t_i)$ soit plus petit que $1/\hat{G}(t)$. Les individus censurés dont le temps de censure est antérieur à t contribuent indirectement à $\hat{G}(t)$. De cette façon, le score de Brier BS peut se rapporter à l'erreur quadratique moyenne entre la fonction de survie estimée et la survie du test pondérée par l'IPCW [127].
Il convient de noter qu'une probabilité prédite de 0.5 pour tous les individus donne une BS de 0.25.
Le score de brier intégré (IBS) permet enfin de mesurer la BS du modèle à tout moment :

$$IBS = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} BS(t)dt$$

Nous avons calculé l'IBS pour la comparaison des 3 méthodes sur le jeu de données réelles NSBCD, la figure 4.14 montre que la validation croisée à 5 blocs améliore nettement la performance, en diminuant très fortement l'IBS par rapport à celle à 2 blocs quelle que soit la méthode.

D'autres schémas de simulation

Variabilité du paramètre de la loi de Weibull et du taux de censure Nos méthodes de simulation nous permettront facilement de fixer ces deux paramètres τ et θ en calculant le paramètre de la loi

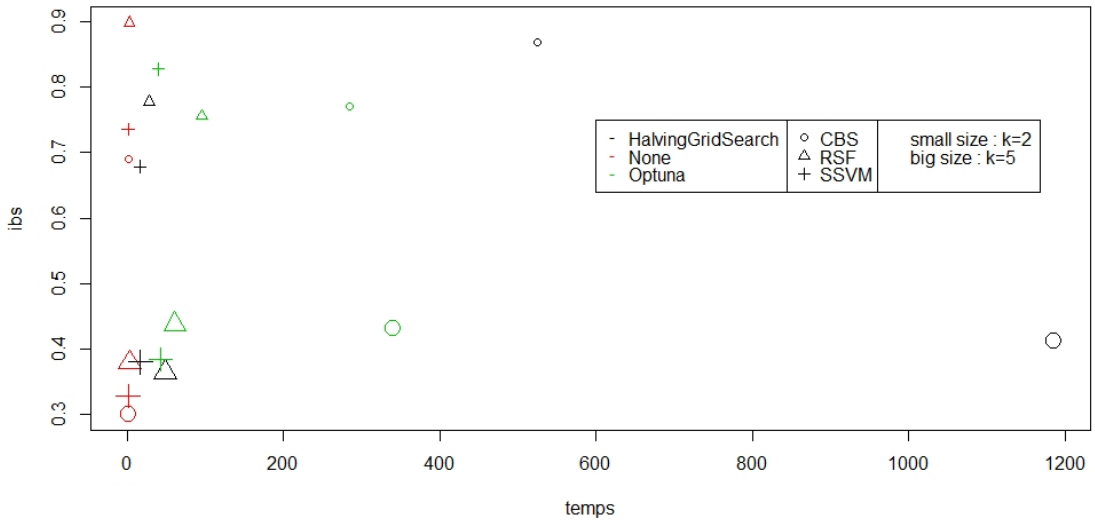


FIGURE 4.14 – Performance en termes d'IBS des méthodes CoxBoost, Random Survival Forest, et Survival SVM par validation croisée 2 et 5 blocs et optimisation ou non des hyperparamètres sur le jeu de données réelles NSBCD.

uniforme qui régit la censure, et ainsi comparer les performances des méthodes en optimisant les hyperparamètres de chacune d'entre elles. La figure 4.15 qui compare des analyses avec les hyperparamètres standards laisse penser que certaines méthodes peuvent pâtir de l'allure de la courbe de survie, qui bien sûr, est très variable sur des données réelles.

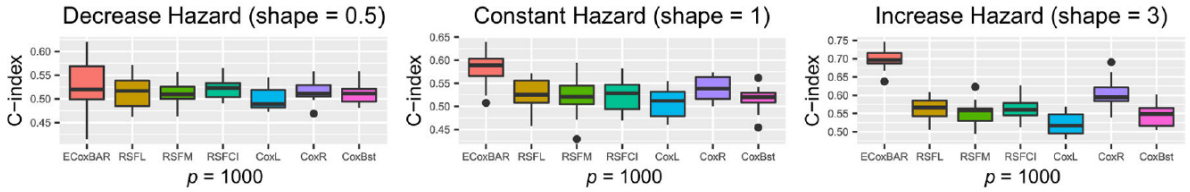


FIGURE 4.15 – Extrait de la FIGURE 1. de l'article [127]. Performance prédictive en termes de C-index pour une fonction de survie à hasard décroissant, constant (Weibull(1) donc une exponentielle), et croissant.

Simulation avec Cmix La méthode de simulation C-mix [173] a été développée avec une distribution de censure de loi géométrique mais peut facilement être modifiée pour une loi uniforme avec nouveau calcul du paramètre θ. Soit $\pi_0 \in [0,1]$ la proportion de patients à faible risque. Est générée une variable latente $Z_i \sim \mathscr{B}(\pi_\beta(x_i))$ où $\pi_\beta(x_i)$ est la probabilité conditionnelle qu'un patient appartienne au groupe à haut risque :

$$\pi_\beta(x) = \frac{\exp({}^t x\beta)}{1+\exp({}^t x\beta)}$$

La distribution conditionnelle au groupe (haut ou faible risque) est une distribution géométrique avec $T_i \sim \mathscr{G}(\alpha_{Z_i})$. Simuler avec une variable latente, qui peut représenter une réalité intéressante, permettra

aussi de fournir des comparaisons plus robustes.

Comparaison des 7 méthodes utilisant Cox

Notamment sur le critère d'importance de variables, nous avons vu que les méthodes (Cox Boost que nous avons utilisée, mais aussi Cox Lasso) qui affectent zéro à de nombreux coefficients peut compliquer notre analyse. Il est par ailleurs connu que Ridge et donc ElasticNet peuvent être plus efficaces en cas de très fortes corrélations des covariables. Cox Boost basée sur les arbres (qualifiée de méthode hybride entre statistique et apprentissage dans la taxonomie décrite précédemment en figure 4.4) est sans doute prometteuse. Enfin Cmix avec la variable latente peut peut-être être moins sensible au bruit. Bien sûr, simuler sous les modèles différents, comme évoqué précédemment, aura donc ici, d'autant plus d'intérêt.

Comparaison de la méthode Cox la plus efficace avec RSF et SSVM

Nous avons appliqué les forêts aléatoires spécifiques à la survie et les Survival SVM à un jeu de données réelles dans la partie 4.2.1, bien évidemment cette analyse avait pour objectif de se familiariser aux méthodes pour les intégrer à notre comparaison de méthodes. Nous avons vu précédemment tout l'enjeu de l'optimisation des hyperparamètres dans les études de simulations où de nombreux papiers présentent des C de Harrell en dessous ou autour de 0.5.

Données réelles

Nous avons commencé à regarder quatre jeux de données réelles disponibles.

— Le jeu de données **DrAsGiven** contient les profils d'expression génique d'échantillons de cancer de l'ovaire traités au Duke University Medical Center et au H. Lee Moffitt Cancer Center and Research Institute. Dans les 117 échantillons utilisés dans cette étude, 22115 caractéristiques géniques et 7 covariables cliniques sont fournies. Les données peuvent être obtenues à partir du package R "dressCheck" de "Bioconductor".

— Le jeu de données **EMTAB386** comprend la signature de l'expression génique de l'ARNm angiogénique et de l'ARNm micro sur 107 cancers ovariens séreux de haut grade à un stade avancé. Après prétraitement, chaque dossier de patient contient 10 357 caractéristiques génétiques et 7 covariables cliniques. Les données peuvent être obtenues à partir du package R "curatedOvarianData".

— Le jeu de données **LungBeer** contient 86 patients atteints d'adénocarcinome pulmonaire. Un total de 7131 covariables est disponible dans l'ensemble de données. Ce jeu de données pouvait être téléchargé sur http ://user.it.uu.se/ liuya610/ dans les années qui ont suivi la publication de l'article [127] et sont maintenant disponibles sur https ://user.it.uu.se/ kripe367/survlab/download.html.

— Le jeu de données **NSBCD**, déjà décrit en partie 4.2.4, comprend les profils d'expression de 549 gènes "intrinsèques" de 115 tumeurs malignes du sein. Ce jeu de données était également disponible à partir du même lien que le précédent.

Les taux de censure diffèrent, l'allure de l'évolution des risques en fonction du temps également. La figure 4.16 résume les caractéristiques de ces 4 jeux de données.

	NSBCD	Lung-Beer	EMTAB386	DrAsGiven
subjects number (n)	115	86	129	119
variables number (p)	549	7129	10357	22115
censorship rate	0,66	0,72	0,57	0,42

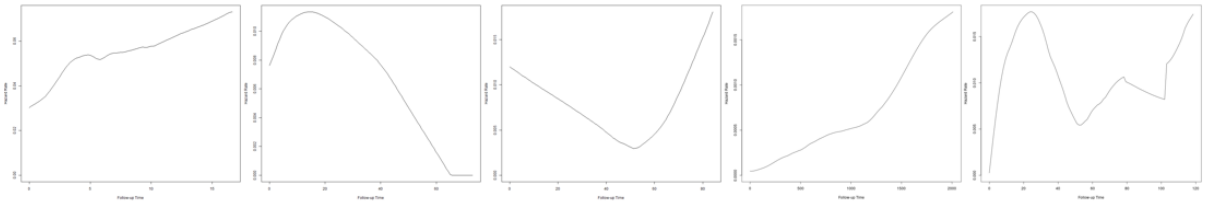


FIGURE 4.16 – Caractéristiques des quatre jeux de données réelles.

4.4.2 Prédictions d'événements récurrents en grande dimension

Dans le cadre de la thèse de Juliette Murris, deux travaux sont prévus à partir du plan de thèse suivant :

- une application des méthodes utilisées dans le premier papier sur des données réelles pour un papier clinique en cancérologie,
- le développement d'une nouvelle méthode basée sur du Boosting Gradient Component Wise adapté au modèle d'Andersen-Gill,
- la définition des scénarios sans doute proches de ceux explorés dans le travail précédent de "Comparaison de méthodes d'analyse de survie en grande dimension", et
- la comparaison de toutes les méthodes implémentées.

Ma contribution à la modélisation des données de survenue d'événement(s) dans un cadre de grande dimension

Les questions cliniques des premières publications de ce chapitre en cancérologie [3, 4, 7, 10] concernaient, entre autres, le problème de dichotomisation d'une variable explicative, optimale pour expliquer la survie des patients. Il n'y avait pas de recommandation méthodologique à ce problème qui par ailleurs, en plus de la question mathématique, soulève également des questions méthodologiques fondamentales de type de modélisation à réaliser et de critères à minimiser. Je suis également revenue sur cette question de critère à minimiser en rajoutant une complexité avec un grand nombre de covariables, revenant ainsi au domaine des données massives en santé et de l'apprentissage statistique. Les nombeuses questions méthodologiques fondamentales qui en découlent ont été abordées par simulations, dont les schémas de simulation complexes étaient basés sur des calculs de densités conjointes pour fixer des paramètres de censure [14]. Ces questions débouchent aujourd'hui sur des projets de recherche d'estimation de biais et l'estimation de certains critères.
Enfin, l'encadrement de la thèse de Juliette Murris a permis de soumettre un premier papier qui a été accepté début janvier 2023 sous réserve de modifications mineures dans Biostatistics and epidemiolgy [16]. Ce travail répertorie les modèles utilisés pour étudier les évènements récurrents en grande dimension, et compare les performances prédictives par simulations.

Chapitre 5

Conclusion

5.1 Synthèse des résultats cliniques ou épidémiologiques

Les 3 premiers articles de la thèse de Canelle Poirier se sont concentrés sur la **grippe, un enjeu de santé publique** de par l'annualité de ses épidémies, imprévisibles sur leur intensité et le moment où elles débutent. Chaque année, dans le monde, jusqu'à 5 millions de cas graves et 500 000 décès peuvent être observés, et les arrêts de travail perturbent l'équilibre du fonctionnement de notre société. Nos travaux ont permis de montrer que les données massives hospitalières étaient très corrélées aux signaux des réseaux de surveillance traditionnelle [11] ; puis qu'elles permettent d'obtenir des estimations en temps réel plus précises que les données du web, ceci comble un peu le délai de consolidation des données de surveillance grâce aux données hospitalières en temps réel [12] ; et enfin que toutes les sources de données (hospitalières, Google, Twitter, et climatiques) apportent de l'information à la prédiction jusqu'à 2 semaines des taux d'incidence grippale de toutes les régions françaises [13].

Le 4$^{\text{ème}}$ article (accepté pour publication début décembre 2022 dans *JMIR Public Health and Surveillance*) s'intéressait à la **gastro-entérite, dont la surveillance est également nécessaire**. La forte augmentation de visites chez les médecins généralistes et dans les services d'urgences hospitalières, perturbe chaque année l'organisation des systèmes de santé. Au niveau régional, les données historiques suffisent pour les prévisions à court terme. Cependant, pour les prévisions à plus long terme, l'inclusion de sources de données externes (Google et EHR) améliore les estimations. Pour la prédiction en temps réel, la performance des modèles est plus grande pour la grippe que pour la gastro-entérite, ce qui n'est plus vrai pour les prédictions à long terme, sauf au niveau régional avec la méthode Random Forest [15].

J'insisterai peu sur les résultats cliniques et épidémiologiques des 6 articles sur lesquels repose le chapitre 3, qui sert surtout à décrire mon expérience en modélisation compartimentale et en modélisation non linéaire à effets mixtes débouchant sur l'approche originale de réunir les deux dans mes projets, en cours et futurs, en propagation d'épidémie. Les deux premiers travaux étaient des articles de ma thèse étudiant la **co-circulation des virus grippaux** [1, 2].

Le premier modèle non linéaire à effets mixtes que j'ai construit pour étudier l'effet de 2 médicaments et leur éventuelle interaction sur la **croissance tumorale**, nous a permis de mettre en évidence une vitesse de croissance ralentie avec telmisartan, et le meilleur modèle après sélection de covariables sur les différents paramètres laisse entrevoir une tendance, cependant non significative (p=0.0687), à une diminution de la taille « finale » quand on ajoute du sunitinib au telmisartan [5].

Les deux articles sur les données de l'étude HEPMEN étaient basés sur l'observation des concentrations en **fer et hepcidine pendant le cycle menstruel** de 90 femmes en bonne santé. Nous avons pu montrer que le fer a stimulé la libération d'hepcidine. Plusieurs covariables, dont la contraception, la quantité de sang perdue et la ferritine, ont influé sur les paramètres. Ce modèle conjoint a apporté une compréhension fondamentale du phénomène [8, 6].

Le dernier article [9] décrit dans le chapitre 3, est la description du **package saemix**, qui a été appliqué sur des données publiques et pédagogiques, mais qui est aujourd'hui utilisé par d'autres équipes sur des problématiques cliniques actuelles telles que, par exemple, la pharmacocinétique de population de la gentamicine chez les patients en pédiatrie à partir d'une étude prospective avec analyse des données à l'aide de saemix [174].

Ces travaux m'ont conduite à tenter une approche non linéaire à effets mixtes basée sur un modèle SIR pour estimer les paramètres moyens et leur variabilité en extrayant d'une série temporelle les épidémies pour les traiter comme des individus statistiques. Cette analyse m'a permis de définir un ensemble de projets de recherche qui en découlent, dont une application sur la **COVID19**.

Dans le chapitre 4, nous quittons le thème de la propagation d'épidémie pour un thème de recherche clinique basé sur la survie et la survenue d'événements. Il démarre par 4 articles en cancérologie, dans le but de proposer aux laboratoires la **détermination du statut de méthylation MGMT pour les patients atteints d'un glioblastome**, en tentant d'identifier la technique d'analyse qui réunirait les meilleures propriétés quant à la valeur prédictive, le coût, la praticabilité et la reproductibilité, à partir d'une cohorte homogène de patients (étude ECOM : essai multicentrique français). Dans le premier article, nous avons donc comparé, entre autres, les performances analytiques et les valeurs prédictives de 5 techniques dont deux quantitatives (pyroséquençage de 5 sites CpG et immunohistochimie) et trois qualitatives (MS-PCR, methylation-sensitive high-resolution melting (MS-HRM) et MethyLight) pour l'analyse MGMT chez 100 patients atteints d'un glioblastome traités par temozolomide. Le pyroséquençage est la meilleure méthode parmi les 5 techniques testées dans cette étude. Cet article a été cité 104 fois (requête pubmed en date du 22/01/23) [3]. L'impact est moindre pour les trois articles qui ont suivi dans la même lignée, mais l'apport sur la recherche de seuils de biomarqueurs dans ce domaine était très attendu [4, 7, 10].

Les technologies modernes permettent de générer des données sur des milliers de variables ou d'observations, selon **la génomique, la radiomique, les bases de données médico-administratives, la surveillance des maladies par des dispositifs médicaux intelligents, etc**. Alors que les données massives décrivent un grand nombre d'observations, les données de grande dimension sont caractérisées par un nombre de variables étudiées p supérieur au nombre d'individus n. De nombreux problèmes méthodologiques doivent alors être étudiés et des recommandations pratiques et pédagogiques sont nécessaires dans le domaine de l'**analyse de survie en grande dimension**. Le travail présenté en 6 pages publiées dans les actes des journées de la statistique de la SFDS, apporte une première réponse en termes d'optimisation des hyperparamètres, de mesures d'évaluation et de validation [14].

Le premier article de thèse de Juliette Murris [16] est la première étude qui s'intéresse à comparer des méthodes standard, des algorithmes de sélection de variables et un réseau neuronal profond pour modéliser des **événements récurrents dans un cadre de grande dimension**. Les progrès des soins médicaux conduisent à l'utilisation de technologies d'intelligence artificielle (IA) embarquées. Le marché en plein essor des dispositifs médicaux d'IA en est une illustration. Ces systèmes sont généralement conçus pour prévenir l'apparition d'événements à l'hôpital, dans une maison de retraite, en consultation externe,

etc. Si ces événements sont susceptibles de se produire de manière répétée et si toutes les données sont disponibles, il est crucial de procéder à une analyse approfondie, robuste et appropriée des événements récurrents. La revue de la littérature a été réalisée dans les règles de l'art en suivant les recommandations méthodologiques, et montre que de nombreuses études cliniques ne s'intéressent au final qu'au premier événement et non à l'ensemble des événements. Nous avons souligné la nécessité actuelle de développer de nouvelles approches, mais aussi d'évaluer leurs performances de manière pertinente, et avons réalisé une comparaison des méthodes disponibles par une étude de simulation.

5.2 Synthèse des résultats de mathématiques appliquées

5.2.1 Développement de modèles complexes pour répondre à une question clinique

Les modèles compartimentaux ont l'avantage de s'adapter à la complexité d'un phénomène biologique, métabolique ou autres, mais nécessitent de construire et formaliser le système avec des hypothèses pour sa simplification, et de le traduire en équations différentielles. Ainsi, j'ai, entre autres, construit le modèle de l'article [1] pour la cocirculation des virus grippaux avec 9 compartiments et 6 paramètres, et participé à la construction de celui de l'article [6] pour la modélisation conjointe du fer et de l'hepcidine avec élimination accrue par perte de sang pendant les règles. Ces constructions nécessitent :

- avant : un travail de lectures médicales, de discussions avec les experts, d'auto-formation au vocabulaire, de recul pour aider les experts à formuler des hypothèses pour répondre aux questions, mais surtout pour simplifier les problématiques. C'est d'ailleurs la partie la plus compliquée et essentielle de distinguer les hypothèses qui simplifient mais ne changent pas substantiellement les effets prédits, des hypothèses erronées qui font une différence importante, comme le concluait l'article de 2008 [175] avec « Le message que cela suggère aux mathématiciens est que les stratégies de contrôle fondées sur des modèles erronés peuvent être dangereuses. »
- pendant : un travail de comparaison de différents modèles structuraux, de différentes structures de variabilité interindividuelle et résiduelle pour les modèles non linéaires à effets mixtes, et de détermination des covariables expliquant tout ou partie de la variabilité entre les individus. Cette démarche nécessite d'une part de choisir un critère statistique permettant de comparer deux modèles et, d'autre part, de se fixer une stratégie quant à l'ordre des modèles à tester. Lorsque nous avons plusieurs modèles de qualité, nous pouvons utiliser des tests afin de retenir le meilleur.
- après : un travail de diagnostics pour s'assurer que le modèle est adapté et/ou de vérification des hypothèses sur lesquelles il a été construit (contrôle des erreurs d'estimation (SE) calculées à partir de la matrice inverse d'information de Fisher, détection de biais à partir de l'étude des prédictions (de population ou individuelles) en fonction des observations, analyse des résidus, des *Visual Predictive Check* (VPC) et des *normalised prediction distribution errors* (npde)).

5.2.2 Simulations pour étudier un phénomène ou vérifier des propriétés statistiques

Travailler sur des données réelles n'est pas simple car les hypothèses faites sur le phénomène qui les a générées sont inconnues et parfois loin de la réalité. A l'inverse, travailler sur des données simulées permet de maîtriser toutes les hypothèses et la vérité est alors connue, mais une grosse difficulté réside à définir les schémas de simulation qui permettront de répondre à notre question. Pour espérer ne pas s'enliser, quelques principes fondamentaux peuvent être suivis :

— Étape 1. Identifier le problème par des données réelles (ou une question clinique) qui ont (ou a) soulevé un problème méthodologique.
— Étape 2. Formuler le problème en définissant l'objectif général de l'étude, les quelques questions spécifiques à traiter et les limites, ainsi que les mesures de performance et la formulation des hypothèses.
— Étape 3. Réaliser la bibliographie et identifier des données réelles similaires.
— Étape 4. Développer des schémas de l'ensemble des étapes nécessaires à la simulation.
— Étape 5. Traduire ces modèles conceptuels mathématiquement (avec les lois qui régiront les variables aléatoires), puis en codant des fonctions (en R ou Python par exemple). Vérifier que le modèle de simulation fonctionne comme prévu. Les techniques de vérification comprennent les traces, la variation des paramètres d'entrée sur leur plage acceptable et la vérification de la sortie, la substitution de constantes aux variables aléatoires et la vérification manuelle et graphique des résultats.
— Étape 5. Formuler et développer la ou les méthodes d'analyse.
— Étape 6. La (ou les) valider en comparant les performances des méthodes dans des conditions connues. Effectuer des tests d'inférence statistique. Examiner les résultats d'un œil critique et expert.
— Étape 7. Documenter en détail les objectifs, les hypothèses et les variables d'entrée. Documenter le plan expérimental (design de simulation).
— Étape 8. Exécuter les simulations conformément à l'étape 7 ci-dessus.
— Étape 9. Interpréter et présenter les résultats. Construire des représentations graphiques.
— Étape 10. Recommander des approfondissements avec des expériences supplémentaires pour augmenter la précision et réduire le biais des estimateurs, la réalisation d'analyses de sensibilité, etc.

Ainsi pour l'étude [14], de nombreux retours en arrière ont été faits, de par des freins difficilement prévisibles au vue de la littérature. En effet, de nombreuses études de simulation font des choix non justifiés voir non justifiables, mettant en péril la réponse présentée, et de nombreuses nouvelles questions se posent au fur et à mesure des études de simulation. Avec une méthodologie rigoureuse, nous avons montré, par exemple, que le C de Harrell n'était pas biaisé en cas de validation croisée emboîtée même quand la taille d'échantillon est petite. La validation croisée 5 blocs n'améliore que très légèrement la performance malgré un temps de calcul deux à trois fois plus long. Nous montrons également une efficacité de l'optimisation des hyperparamètres, comme espérée, mais rarement réalisée dans les études de simulation à cause de l'augmentation des temps de calcul et du niveau de complexité.

Plus la question est précise et issue d'une problématique clinique concrète au début du travail, plus le schéma de simulations peut être réduit, nous étions dans cette situation dans nos travaux [1] et [16].

5.2.3 Vulgarisation et adaptation des concepts mathématiques sous-jacents

En statistique appliquée, il est nécessaire de s'approprier les développements mathématiques sous-jacents afin de pouvoir les programmer, voire les vulgariser pour l'utilisation des méthodes ou fonctions publiées. Ainsi, l'article [9] décrit le principe de l'algorithme Stochastic Approximation of EM (SAEM) en faisant référence aux preuves mathématiques de convergence. J'ai présenté le package R saemix que nous avons développé à partir de sa version Matlab, aux Rencontres R à Bordeaux en 2012 [90].

Dans l'article [14], nous avons dû adapter les calculs de Wan [147] (produit de fonctions de densité, intégration sur le domaine de définition, etc.) à notre choix de loi normale pour les covariables X, de loi

de Weibull pour les durées de survie T et de loi uniforme pour les censures C, pour obtenir le paramètre θ de la distribution de censure. Ceci nous a permis de fixer les taux de censure à la valeur souhaitée même si les scénarios font varier le nombre de covariables actives ou le paramètre ν de la loi de Weibull.

Enfin, dans l'article [1], j'ai dû étudier la stabilité des points d'équilibre de notre système à 9 compartiments. Si le point particulier qui représente une population sans aucun infecté n'est pas stable, l'introduction d'un seul infecté peut faire démarrer l'épidémie. Pour qu'un point stationnaire soit localement asymptotiquement stable, il faut vérifier que les valeurs propres de la Jacobienne, évaluées à ce point, aient leur partie réelle négative. Après calculs analytiques, l'équation obtenue nous a permis d'identifier le nombre de reproduction R0, qui doit être inférieur à 1 pour bloquer l'épidémie.

5.3 Perspectives

Mes perspectives de recherche à court et plus long termes ont été développées en une quinzaine de pages en fin de chapitre 3, 4 et 5. Je vais les résumer ici en quelques lignes :

Dans le chapitre 2, l'approche d'apprentissage profond (LSTM) a été testée rapidement et semble prometteuse. Un sujet de recherche à part entière, peut être envisagé, pour adapter un réseau neuronal complexe qui inclura les sources de données externes telles que les données Google ou Twitter, pour permettre la prévision à long terme des incidences de différents syndrômes tels que la grippe ou la gastroentérite. Une autre perspective intéressante est la modélisation conjointe SIR (Susceptible-Infectious-Removed) pour la période épidémique, et apprentissage statistique le reste du temps.

J'ai présenté dans le chapitre 3 mon approche non linéaire à effets mixtes basée sur un modèle SIR pour estimer les paramètres moyens et leur variabilité en extrayant d'une série temporelle les épidémies pour les traiter comme des individus statistiques. Ce travail préliminaire, qui a soulevé plusieurs freins, débouche ainsi sur un ensemble de projets de recherche. Pour l'intérêt épidémiologique, l'analyse aurait dû prendre en compte une structure d'âge, j'ai obtenu les données du Réseau Sentinelles, j'ai complexifié mon modèle avec une matrice de contacts entre les classes d'âge, et malgré l'apport de données, le nombre de paramètres à estimer dans ce modèle très alourdi, a conduit à une modélisation qui n'était plus satisfaisante. J'ai alors préféré mettre en suspend ce travail pour lancer une collaboration anglaise en revenant à la base de l'étude d'identifiabilité des modèles SIR.

Nous travaillons donc actuellement avec Dave Woods de l'Institut de Recherche des Sciences Statistiques de Southampton (S3RI) de l'Université de Southampton, et Samuel Jackson du Département de Sciences Mathématiques de l'Université de Durham, qui a été reçu 2 semaines à l'IRMAR sur financements per diem pour travailler sur ce projet avec moi en septembre 2022. Ce travail théorique et de simulations nous permettra de proposer des recommandations pratiques et pédagogiques pour diagnostiquer facilement l'identifiabilité des paramètres avant toute utilisation de modèles compartimentaux. La pandémie de coronavirus (COVID-19) a considérablement accru la sensibilisation du public et son appréciation de l'utilité des modèles dynamiques. Mais la diffusion de prévisions contradictoires des modèles a mis en évidence leurs limites. Un article récent [113] a passé en revue les différents modèles proposés dans la littérature les premiers mois de la pandémie, avec 36 structures de modèles et en évaluant leur capacité à fournir des informations fiables pour différentes situations. Cette étude approfondie montre par exemple que le taux de transmission est identifiable seulement pour 59/98. Nous utiliserons trois modèles pour appliquer nos recommandations d'outils diagnostiques, et nous avons regardé la faisabilité d'évaluation de ces modèles sur les données réelles de la $2^{\text{ème}}$ vague de COVID19 au Royaume Uni et en France.

Quand cette partie sera réalisée je pourrai reprendre le travail de modélisation non linéaire à effets mixtes avec un modèle SIR pour étudier les épidémies saisonnières françaises de grippe, voire proposer un sujet de doctorat. J'envisage aussi un objectif prédictif (prédiction dynamique individuelle) au travers d'une collaboration avec Solène Desmée, maîtresse de conférences à Tours dans l'unité Inserm SPHERE. Ceci pourrait nous permettre de répondre à plusieurs questions du type :

- A partir de quand les paramètres individuels deviennent stables ?
- Peut-on définir un indicateur de stabilité, qui permettrait de savoir prédire correctement à 3 semaines, le moment du pic et la taille de l'épidémie ?
- L'étude des épidémies saisonnières peut-elle permettre la prédiction de la pandémie de grippe H1N1 de 2009 (en fixant p_{S_0} à 100%-$p_{vaccination}$) ?

J'ai présenté dans le chapitre 4 les résultats publiés en 6 pages dans les actes des journées de la statistique de la SFDS [14], qui ne sont bien évidemment que la première partie d'un papier qui pourrait s'intituler "Les risques de biais des critères d'évaluation des performances en analyse de survie en grande dimension sur petits échantillons". Nous envisageons d'étudier d'autres critères d'évaluation (le C sur échantillon reconstruit, le C de Uno, et le Score de Brier intégré (IBS)), mais aussi d'autres schémas de simulation (en faisant varier le paramètre de la loi de Weibull et le taux de censure, et en simulant avec Cmix par exemple). Nous comparerons différentes méthodes statistiques et d'apprentissage pour l'analyse des données simulées (différents Cox, RSF et SSVM). Enfin, les résultats des simulations pourront être confrontés à l'application sur des jeux de données réelles que nous avons déjà commencé à regarder. Tout ceci pourrait, là aussi, faire l'objet d'une proposition de thèse.

Enfin, dans le cadre de la thèse de Juliette Murris, deux travaux supplémentaires sont prévus à partir du plan de thèse suivant :

- une application des méthodes utilisées dans le premier papier sur des données réelles pour un papier clinique en cancérologie,
- le développement d'une nouvelle méthode basée sur du Boosting Gradient Component Wise adapté au modèle d'Andersen-Gill,
- la définition des scénarios sans doute proches de ceux explorés dans le travail précédent de "Comparaison de méthodes d'analyse de survie en grande dimension", et
- la comparaison de toutes les méthodes implémentées.

Pour conclure

J'ai eu de nombreuses occasions de travailler sur des domaines très variés. Je n'ai cherché à faire des liens dans ce mémoire qu'avec ceux qui correspondent à des encadrements de thèses et/ou aux directions que je souhaite porter à l'avenir. Ces travaux reflètent également une direction plus affirmée au fil du temps vers des questions plus méthodologiques, bien que toujours générées par des questions cliniques ou épidémiologiques. Mon affiliation à l'IRMAR et mon émargement au CIC me permettent de travailler avec des collègues des deux laboratoires. La rédaction de ce mémoire m'a permis de mieux prendre conscience de mes compétences transdisciplinaires, et de plus les voir aujourd'hui comme un atout et une expertise spécifique.

Bibliographie

[1] Lavenu A, Valleron AJ, Carrat F. Exploring cross-protection between influenza strains by an epidemiological model. *Virus Res* Jul 2004 ; **103**(1-2) :101–105.

[2] Carrat F, Lavenu A. Heterosubtypic immunity to influenza : right hypothesis, wrong comparison. *J Infect Dis* Jun 2006 ; **193**(11) :1613–1614.

[3] Quillien V, Lavenu A, Karayan-Tapon L, Carpentier C, Labussière M, Lesimple T, Chinot O, Wager M, Honnorat J, Saikali S, *et al.*. Comparative assessment of 5 methods (methylation-specific polymerase chain reaction, MethyLight, pyrosequencing, methylation-sensitive high-resolution melting, and immunohistochemistry) to analyze O6-methylguanine-DNA-methyltranferase in a series of 100 glioblastoma patients. *Cancer* Sep 2012 ; **118**(17) :4201–4211.

[4] Quillien V, Lavenu A, Sanson M, Legrain M, Dubus P, Karayan-Tapon L, Mosser J, Ichimura K, Figarella-Branger D. Outcome-based determination of optimal pyrosequencing assay for MGMT methylation detection in glioblastoma patients. *J Neurooncol* Feb 2014 ; **116**(3) :487–496.

[5] Verhoest G, Dolley-Hitze T, Jouan F, Belaud-Rotureau MA, Oger E, Lavenu A, Bensalah K, Arlot-Bonnemains Y, Collet N, Rioux-Leclercq N, *et al.*. Sunitinib combined with angiotensin-2 type-1 receptor antagonists induces more necrosis : a murine xenograft model of renal cell carcinoma. *Biomed Res Int* 2014 ; **2014** :901371.

[6] Angeli A, Lainé F, Lavenu A, Ropert M, Lacut K, Gissot V, Sacher-Huvelin S, Jezequel C, Moignet A, Laviolle B, *et al.*. Joint Model of Iron and Hepcidin During the Menstrual Cycle in Healthy Women. *AAPS J* Mar 2016 ; **18**(2) :490–504.

[7] Quillien V, Lavenu A, Ducray F, Joly MO, Chinot O, Fina F, Sanson M, Carpentier C, Karayan-Tapon L, Rivet P, *et al.*. Validation of the high-performance of pyrosequencing for clinical MGMT testing on a cohort of glioblastoma patients from a prospective dedicated multicentric trial. *Oncotarget* 09 2016 ; **7**(38) :61916–61929.

[8] Lainé F, Angeli A, Ropert M, Jezequel C, Bardou-Jacquet E, Deugnier Y, Gissot V, Lacut K, Sacher-Huvelin S, Lavenu A, *et al.*. Variations of hepcidin and iron-status parameters during the menstrual cycle in healthy women. *Br J Haematol* 12 2016 ; **175**(5) :980–982.

[9] Comets E, Lavenu A, Lavielle M. Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software* 2017 ; **80**(3) :1–41, doi :10.18637/jss.v080.i03. URL `https://www.jstatsoft.org/index.php/jss/article/view/v080i03`.

[10] Quillien V, Lavenu A, Ducray F, Meyronet D, Chinot O, Fina F, Sanson M, Carpentier C, Karayan-Tapon L, Rivet P, *et al.*. Clinical validation of the CE-IVD marked Therascreen MGMT kit in a cohort of glioblastoma patients. *Cancer Biomark* Dec 2017 ; **20**(4) :435–441.

[11] Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, Chazard E, Lavenu A, Cuggia M. Leveraging hospital big data to monitor flu epidemics. *Comput Methods Programs Biomed* Feb 2018 ; **154** :153–160.

[12] Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods : Comparison Study. *JMIR Public Health Surveill* Dec 2018 ; **4**(4) :e11361.

[13] Poirier C, Hswen Y, Bouzillé G, Cuggia M, Lavenu A, Brownstein JS, Brewer T, Santillana M. Influenza forecasting for French regions combining EHR, web and climatic data sources with a machine learning ensemble approach. *PLoS One* 2021 ; **16**(5) :e0250890.

[14] Lavenu A, Murris J, Mareau A, Rouzé T, Fromont M, Gares V, Katsahian S. Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation. *53ème Journées de Statistique de la SFdS, p879-885. 13-17 juin 2022, Lyon, France*, 2022.

[15] Poirier C, Bouzillé G, Bertaud V, Cuggia M, , Santillana M, Lavenu A. Gastroenteritis forecasting assessing the use of web and EHR data with machine learning approaches. *JMIR Public Health and Surveillance* accepté fin 2022 ; .

[16] Murris J, Charles-Nelson A, Lavenu A, Katsahian S. Towards Filling the Gaps around Recurrent Events in High Dimensional Framework : Literature Review and Early Comparison. *Biostatistics and Epidemiology* accepté avec révisions mineures début janvier 2023 ; .

[17] Nevoret C, Tran Y, Guendouz S, Lavenu A, Damy T, Katsahian S, Tropeano A. Healthcare trajectories and mortality in heart failure. *European journal of heart failure* soumis ; .

[18] Azim A, Freeman A, Lavenu A, Mistry H, Haitchi HM, Newell C, Cheng Y, Thirlwall Y, Harvey M, Barber C, *et al.*. New perspectives on difficult asthma ; sex and age of asthma-onset based phenotypes. *The Journal of Allergy and Clinical Immunology : In Practice* 06 2020 ; **8**, doi :10.1016/j.jaip.2020.05.053.

[19] Qu Y, Zhang R, Cui P, Song G, Duan Z, Lei F. Evolutionary genomics of the pandemic 2009 h1n1 influenza viruses (ph1n 1v). *Virology journal* 05 2011 ; **8** :250, doi :10.1186/1743-422X-8-250.

[20] Zhu X, Zhou X, Zhang Y, Sun X, Liu H, Zhang Y. Reporting and methodological quality of survival analysis in articles published in chinese oncology journals. *Medicine* 12 2017 ; **96** :e9204, doi :10.1097/MD.0000000000009204.

[21] Bourany T. Les 5v du big data. *Regards croisés sur l'économie* 01 2018 ; **n°23** :27, doi :10.3917/rce.023.0027.

[22] Smith H, Sweeting M, Morris T, Crowther M. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagnostic and Prognostic Research* 06 2022 ; **6** :10, doi :10.1186/s41512-022-00124-y.

[23] Saporta G. Expliquer ou prédire ? les nouveaux défis. *Chimiometrie 2017, Jan 2017, Paris, France*, 2017.

[24] Breiman L. Statistical modeling : The two cultures (with comments and a rejoinder by the author). *Statistical Science* 08 2001 ; **16**, doi :10.1214/ss/1009213726.

[25] Shmueli G. To explain or to predict ? *Statistical Science* 01 2011 ; **25**, doi :10.1214/10-STS330.

[26] Bühlmann P. Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research* 06 2013 ; **77**, doi :10.1007/s00186-012-0404-7.

[27] Shiffrin R. Drawing causal inference from big data. *Proceedings of the National Academy of Sciences* 07 2016 ; **113** :7308–7309, doi :10.1073/pnas.1608845113.

[28] Bousquet C, Beltramin D. *Machine Learning in Medicine : To Explain, or Not to Explain, That Is the Question*, vol. 294. 2022, doi :10.3233/SHTI220407.

[29] Formenty P, Roth C, Gonzalez-Martin F, Grein T, Ryan M, Drury P, Kindhauser M, Rodier G. Emergent pathogens, international surveillance and international health regulations (2005)]. *Médecine et maladies infectieuses* 02 2006 ; **36** :9–15.

[30] SPF. Sante publique france - qui sommes-nous. `https://www.santepubliquefrance.fr/Sante-publique-France/Qui-sommes-nous/` 2022. Accessed : 2022-07-01.

[31] RS. Reseau sentinelles - accueil. `https://websenti.u707.jussieu.fr/sentiweb/` 2022. Accessed : 2022-07-01.

[32] Josseran L, Fouillet A. La surveillance syndromique : bilan et perspective d'un concept prometteur. *Revue D Epidemiologie Et De Sante Publique* 2013 ; **61** :163–170.

[33] Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A* Nov 2015 ; **112**(47) :14 473–14 478.

[34] Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, Brownstein JS. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. *Sci Rep* 05 2016 ; **6** :25 732.

[35] Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci U S A* Nov 2015 ; **112**(47) :14 473–14 478.

[36] Serfling R. Methods for current statistical analysis of excess pneumonia-in- fluenza deaths. *Public Health Rep* 1963 ; **78** :494–506.

[37] Pelat C, Boëlle PY, Cowling BJ, Carrat F, Flahault A, Ansart S, Valleron AJ. Online detection and quantification of epidemics. *BMC Med Inform Decis Mak* Oct 2007 ; **7** :29.

[38] GT. Google trends - france. `https://trends.google.fr/trends/?geo=FR/` 2022. Accessed : 2022-07-01.

[39] Team RC. R : A language and environment for statistical computing. *MSOR connections* 2014 ; **1**.

[40] Massicotte P, Eddelbuettel D. R functions to perform and display google trends queries 03 2016 ; .

[41] Lowen AC, Steel J. Roles of humidity and temperature in shaping influenza seasonality. *J Virol* Jul 2014 ; **88**(14) :7692–7695.

[42] Lowen AC, Mubareka S, Steel J, Palese P. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog* Oct 2007 ; **3**(10) :1470–1476.

[43] Tamerius JD, Shaman J, Alonso WJ, Alonso WJ, Bloom-Feshbach K, Uejio CK, Comrie A, Viboud C. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathog* Mar 2013 ; **9**(3) :e1003 194.

[44] Lawrence MG. The relationship between relative humidity and the dewpoint temperature in moist air : A simple conversion and applications. *Bulletin of the American Meteorological Society* 2005 ; **86**(2) :225 – 234, doi :10.1175/BAMS-86-2-225. URL `https://journals.ametsoc.org/view/journals/bams/86/2/bams-86-2-225.xml`.

[45] Zou H, Hastie T. Zou h, hastie t. regularization and variable selection via the elastic net. j r statist soc b. 2005 ;67(2) :301-20. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 04 2005 ; **67** :301 – 320, doi :10.1111/j.1467-9868.2005.00503.x.

[46] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)* 01 1996 ; **58** :267–288, doi :10.1111/j.2517-6161.1996.tb02080.x.

[47] Hoerl A, Kennard R. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics* 01 1970 ; **8** :27–51.

[48] Trevor Hastie JF Robert Tibshirani. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer Series in Statistics : Stanford, 2008.

[49] RCT. R : A language and environment for statistical computing. `https://www.R-project.org/` 2022. R Foundation for Statistical Computing, Accessed : 2022-07-01.

[50] Tibshirani R, Hastie T, Friedman J. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 02 2010 ; **33**, doi :10.1163/ej.9789004178922.i-328.7.

[51] Breiman L. Random forests. *Machine Learning* 2004 ; **45** :5–32.

[52] Fernández-Delgado M, Cernadas E, Barro S, Amorim DG. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research* 2014 ; **15** :3133–3181.

[53] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest : a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003 ; **43**(6) :1947–1958.

[54] Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. Random forests for classification in ecology. *Ecology* Nov 2007 ; **88**(11) :2783–2792.

[55] Shotton J, Sharp T, Kipman A, Fitzgibbon AW, Finocchio M, Blake A, Cook M, Moore R. Real-time human pose recognition in parts from single depth images. *CVPR 2011* 2011 ; :1297–1304.

[56] Kane MJ, Price N, Scotch M, Rabinowitz PM. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics* 2012 ; **15**.

[57] Dudek G. Short-term load forecasting using random forests. *IEEE Conf. on Intelligent Systems*, 2014.

[58] Lahouar A, Slama JBH. Random forests model for one day ahead load forecasting. *IREC2015 The Sixth International Renewable Energy Congress* 2015 ; :1–6.

[59] Liaw A, Wiener M. Classification and regression by randomforest. *Forest* 11 2001 ; **23**.

[60] Goehry B, Yan H, Goude Y, Massart P, Poggi JM. Random forests for time series : Accepted - november 2021. *REVSTAT-Statistical Journal* Nov 2021 ; URL `https://revstat.ine.pt/index.php/REVSTAT/article/view/400`.

[61] Vapnik V. *The Nature of Statistical Learning Theory*, vol. 8. 2000 ; 1–15, doi :10.1007/978-1-4757-3264-1_1.

[62] Khan F, Enzmann F, Kersten M. Multi-phase classification by a least-squares support vector machine approach in tomography images of geological samples. *Solid Earth* 03 2016 ; **7** :481–492, doi :10.5194/se-7-481-2016.

[63] Moradzadeh A, Mansour Saatloo A, Mohammadi-ivatloo B, Anvari-Moghaddam A. Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings. *Applied Sciences* 05 2020 ; **10**, doi :10.3390/app10113829.

[64] Zhang A, Lipton Z, Li M, Smola A. *Dive into Deep Learning.* 2021.

[65] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation* 12 1997 ; **9** :1735–80, doi :10.1162/neco.1997.9.8.1735.

[66] Lu F, Hattab M, Clemente C, Biggerstaff M, Santillana M. Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches. *Nature Communications* 01 2019 ; **10**, doi :10.1038/s41467-018-08082-0.

[67] Farthing M. Diarrhoea : A significant worldwide problem. *International journal of antimicrobial agents* 03 2000 ; **14** :65–9, doi :10.1016/S0924-8579(99)00149-1.

[68] Kosek M, Bern C, Guerrant R. The global burden of diarrheal disease, as estimated from studies published between 1992 and 2000. *Bulletin of the World Health Organization* 02 2003 ; **81** :197–204, doi :10.1590/S0042-96862003000300010.

[69] Cauteren D, de Valk H, Vaux S, Le Strat Y, Vaillant V. Burden of acute gastroenteritis and healthcare-seeking behaviour in france : A population-based study. *Epidemiology and infection* 06 2011 ; **140** :697–705, doi :10.1017/S0950268811000999.

[70] Guo T, Lin T, Antulov-Fantulin N. Exploring interpretable lstm neural networks over multi-variable data 05 2019.

[71] E Ette PW. *Pharmacometrics : The Science of Quantitative Pharmacology.* John Wiley & Sons : Hoboken, 2007.

[72] Lee J, Garnett C, Gobburu J, Bhattaram V, Brar S, Earp J, Pravin R, Krudys K, Lesko L, Li F, *et al.*. Impact of pharmacometric analyses on new drug approval and labelling decisions. *Clinical pharmacokinetics* 10 2011 ; **50** :627–35, doi :10.2165/11593210-000000000-00000.

[73] Beal S, Boeckmann L, Bauer R, Sheiner L. Nonmem user's guides. (1989–2009). 2009.

[74] Lindstrom M, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics* 1990 ; **46 3** :673–87.

[75] Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995 ; **4** :12–35.

[76] Comets E, Mentré F. Evaluation of tests based on individual versus population modeling to compare dissolution curves. *Journal of Biopharmaceutical Statistics* 2001 ; **11** :107 – 123.

[77] Bertrand J, Comets E, Mentré F. Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters. *Journal of Biopharmaceutical Statistics* 2008 ; **18** :1084 – 1102.

[78] Molenberghs G, Verbeke G. Models for discrete longitudinal data. 2005.

[79] Savic RM, Mentré F, Lavielle M. Implementation and evaluation of the saem algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics–pharmacodynamics. *The AAPS Journal* 2010 ; **13** :44–53.

[80] Plan EL, Maloney A, Mentré F, Karlsson MO, Bertrand J. Performance comparison of various maximum likelihood nonlinear mixed-effects estimation methods for dose–response models. *The AAPS Journal* 2012 ; **14** :420–432.

[81] Lavielle M. Mixed effects models for the population approach : Models, tasks, methods and tools. 2014.

[82] Bates DM, Machler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 2014 ; **67** :1–48.

[83] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em - algorithm plus discussions on the paper. 1977.

[84] Delyon B, Lavielle M, Moulines É. Convergence of a stochastic approximation version of the em algorithm. *Annals of Statistics* 1999 ; **27** :94–128.

[85] Wolfinger RD. Laplace's approximation for nonlinear mixed models. *Biometrika* 1993 ; **80** :791–795.

[86] Wei GCG, Tanner MA. Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika* 1990 ; **77** :649–652.

[87] The MathWorks Inc. Matlab – the language of technical computing, version r2014b. `http://www.mathworks.com/products/matlab/` 2014. Natick.

[88] Comets E, Lavenu A, Lavielle M. saemix : Stochastic approximation expectation maximization (saem) algorithm. `https://CRAN.R-project.org/package=saemix` 2017.

[89] Brendel K, Comets E, Laffont CM, Laveille C, Mentré F. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. *Pharmaceutical Research* 2006 ; **23** :2036–2049.

[90] Lavenu A, Comets E, Lavielle M. une version r de l'algorithme saem pour l'estimation de paramètres dans les modèles non linéaires à effets mixtes. *1ères Rencontres R, Bordeaux*, 2012.

[91] Comets E, Lavenu A, Lavielle M. saemix, an r version of the saem algorithm. *20th Meeting of the Population Approach Group in Europe, Athens, Greece. Abstract 2173*, 2011.

[92] Kautz L. Rôle de bmp6 et de hfe dans la régulation de l'entrée du fer dans l'organisme. 2009.

[93] Yaari R, Katriel G, Huppert A, Axelsen J, Stone L. Modelling seasonal influenza : The role of weather and punctuated antigenic drift. *Journal of the Royal Society, Interface / the Royal Society* 04 2013 ; **10** :20130 298, doi :10.1098/rsif.2013.0298.

[94] Truscott J, Fraser C, Cauchemez S, Meeyai A, Hinsley W, Donnelly C, Ghani A, Ferguson N. Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal inuenza. *Journal of the Royal Society, Interface / the Royal Society* 06 2011 ; **9** :304–12, doi :10.1098/rsif.2011.0309.

[95] Samanlioglu F, Bilge A. An overview of the 2009 a(h1n1) pandemic in europe : Efficiency of the vaccination and healthcare strategies. *Journal of Healthcare Engineering* 01 2016 ; **2016** :1–13, doi :10.1155/2016/5965836.

[96] Urabe CT, Tanaka G, Aihara K, Mimura M. Parameter scaling for epidemic size in a spatial epidemic model with mobile individuals. *PLOS ONE* 12 2016 ; **11** :e0168 127, doi :10.1371/journal.pone.0168127.

[97] Jackson C, Vynnycky E, Mangtani P. The relationship between school holidays and transmission of influenza in england and wales. *American Jnl of Epidemiology* 10 2016 ; **184**, doi :10.1093/aje/kww083.

[98] Bilge A, Samanlioglu F, Ergonul O. On the uniqueness of epidemic models fitting a normalized curve of removed individuals. *Journal of mathematical biology* 10 2014 ; **71**, doi :10.1007/s00285-014-0838-z.

[99] Boianelli A, Nguyen V, Ebensen T, Schulze K, Wilk E, Sharma N, Stegemann-Koniszewski S, Bruder D, Toapanta F, Guzman CA, *et al.*. Modeling influenza virus infection : A roadmap for influenza research. *Viruses* 10 2015 ; **7** :5274–5304, doi :10.3390/v7102875.

[100] Chowell G, Viboud C, Simonsen L, Moghadas S. Characterizing the reproduction number of epidemics with early subexponential growth dynamics. *Journal of the Royal Society, Interface* 10 2016 ; **13**, doi :10.1098/rsif.2016.0659.

[101] Hickmann K, Fairchild G, Priedhorsky R, Generous N, Hyman J, Deshpande A, Del Valle S. Forecasting the 2013–2014 influenza season using wikipedia. *PLoS computational biology* 10 2014 ; **11**, doi :10.1371/journal.pcbi.1004239.

[102] Anderson RM, May RM. *Infectious diseases of humans : dynamics and control*. Oxford university press, 1991.

[103] Pybus O, Charleston M, Gupta S, Rambaut A, Holmes E, Harvey P. The epidemic behavior of the hepatitis c virus. *Science (New York, N.Y.)* 07 2001 ; **292** :2323–5, doi :10.1126/science.1058321.

[104] Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature* 2005 ; **437**(7056) :209–214.

[105] Ferguson NM, Cummings DA, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. *Nature* 2006 ; **442**(7101) :448–452.

[106] Cauchemez S, Valleron AJ, Boelle PY, Flahault A, Ferguson N. Estimating the impact of school closure on influenza transmission from sentinel data. *Nature* 05 2008 ; **452** :750–4, doi :10.1038/nature06732.

[107] Leroux A, Xiao L, Crainiceanu C, Checkley W. Dynamic prediction in functional concurrent regression with an application to child growth. *Statistics in Medicine* 12 2017 ; **37**, doi :10.1002/sim.7582.

[108] Desmée S, Mentré F, Veyrat-Follet C, Sebastien B, Guedj J. Nonlinear joint models for individual dynamic prediction of risk of death using hamiltonian monte carlo : Application to metastatic prostate cancer. *BMC Medical Research Methodology* 07 2017 ; **17**, doi :10.1186/s12874-017-0382-9.

[109] Tuncer N, Le TT. Structural and practical identifiability analysis of outbreak models. *Math Biosci* 05 2018 ; **299** :1–18.

[110] Evans ND, White LJ, Chapman MJ, Godfrey KR, Chappell MJ. The structural identifiability of the susceptible infected recovered model with seasonal forcing. *Math Biosci* Apr 2005 ; **194**(2) :175–197.

[111] Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* Feb 2009 ; **6**(31) :187–202.

[112] Soetaert K, Petzoldt T. Inverse modelling, sensitivity and monte carlo analysis in r using package fme. *Journal of Statistical Software, Articles* 2010 ; **33**(3).

[113] Massonis G, Banga J, Villaverde A. Structural identifiability and observability of compartmental models of the covid-19 pandemic. *Annual Reviews in Control* 12 2020 ; **51**, doi :10.1016/j.arcontrol.2020.12.001.

[114] Kaushal S, Rajput AS, Bhattacharya S, Vidyasagar M, Kumar A, Prakash MK, Ansumali S. Estimating the herd immunity threshold by accounting for the hidden asymptomatics using a COVID-19 specific model. *PLoS One* 2020 ; **15**(12) :e0242 132.

[115] Ponce M, Sandhel A. covid19.analytics : An r package to obtain, analyze and visualize data from the 2019 coronavirus disease pandemic. *Journal of Open Source Software* 04 2021 ; **6** :2995, doi :10.21105/joss.02995.

[116] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1* 04 1966 ; **50** :163–70.

[117] Peto R, Peto J. Asymptotically efficient rank invariate test procedures (with discussion). *J R Stat Soc [A]* 01 1972 ; **1** :135–185.

[118] Heagerty P, Zheng Y. Survival model predictive accuracy and roc curves. *Biometrics* 04 2005 ; **61** :92–105, doi : 10.1111/j.0006-341X.2005.030814.x.

[119] Bauchet L, Rigau V, Mathieu-Daudé H, Figarella-Branger D, Duffau H, Palusseau L, Bauchet F, Fabbro M, Campello C, Capelle L, *et al.*. French brain tumor data bank : Methodology and first results on 10,000 cases. *Journal of neuro-oncology* 10 2007 ; **84** :189–99, doi :10.1007/s11060-007-9356-9.

[120] Stupp R, Mason W, Bent M, Weller M, Fisher B, Taphoorn M, Belanger K, Brandes A, Marosi C, Bogdahn U, *et al.*. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *The New England journal of medicine* 03 2005 ; **352** :987–96, doi :10.1056/NEJMoa043330.

[121] Hermisson M, Klumpp A, Wick W, Wischhusen J, Nagel G, Roos W, Kaina B, Weller M. Hermisson m, klumpp a, wick w, wischhusen j, nagel g, roos w, kaina b, weller mo6-methylguanine dna methyltransferase and p53 status predict temozolomide sensitivity in human malignant glioma cells. j neurochem 96 : 766-776. *Journal of neurochemistry* 02 2006 ; **96** :766–76, doi :10.1111/j.1471-4159.2005.03583.x.

[122] Stupp R, Hegi M. Methylguanine methyltransferase testing in glioblastoma : when and how ? j clin oncol 25(12) : 1459-1460. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 05 2007 ; **25** :1459–60, doi :10.1200/JCO.2006.09.7139.

[123] Budczies J, Klauschen F, Sinn B, Győrffy B, Schmitt W, Darb-Esfahani S, Denkert C. Cutoff finder : A comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PloS one* 12 2012 ; **7** :e51 862, doi : 10.1371/journal.pone.0051862.

[124] Foucher Y, Giral M, Soulillou JP, Daurès JP. Cut-off estimation and medical decision making based on a continuous prognostic factor : The prediction of kidney graft failure. *The international journal of biostatistics* 01 2012 ; **8**, doi :10.2202/1557-4679.1215.

[125] Heagerty P, Lumley T, Pepe M. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* 06 2000 ; **56** :337–44, doi :10.1111/j.0006-341X.2000.00337.x.

[126] Akritas M. Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics* 09 1994 ; **22**, doi :10.1214/aos/1176325630.

[127] Wang H, Li G. Extreme learning machine Cox model for high-dimensional survival analysis. *Stat Med* 05 2019 ; **38**(12) :2139–2156.

[128] Pittman J, Huang E, Dressman H, Horng CF, Cheng S, Tsou MH, Chen CY, Bild A, Iversen E, Huang A, *et al.*. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America* 07 2004 ; **101** :8431–6, doi : 10.1073/pnas.0401736101.

[129] Vabalas A, Gowen E, Poliakoff E, Casson A. Machine learning algorithm validation with a limited sample size. *PLOS ONE* 11 2019 ; **14** :e0224365, doi :10.1371/journal.pone.0224365.

[130] Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics* 02 2008 ; **9** :14, doi :10.1186/1471-2105-9-14.

[131] Tutz G, Binder H. Boosting ridge regression. *Computational Statistics and Data Analysis* 2007 ; **51**(12) :6044–6059, doi :https://doi.org/10.1016/j.csda.2006.11.041. URL `https://www.sciencedirect.com/science/article/pii/S0167947306004749`.

[132] Binder H, Schumacher M. Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics* 2009 ; **10** :18 – 18.

[133] Leblanc M, Crowley J. Survival trees by goodness of split. *Journal of The American Statistical Association - J AMER STATIST ASSN* 06 1993 ; **88** :457–467, doi :10.1080/01621459.1993.10476296.

[134] Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. *The Annals of Applied Statistics* 12 2008 ; **2**, doi :10.1214/08-AOAS169.

[135] Ishwaran H, Lu M. *Random Survival Forests.* 2019 ; 1–13, doi :10.1002/9781118445112.stat08188.

[136] Aalen O. Nonparametric inference for a family of counting processes. *Ann. Statist.* 07 1978 ; **6**, doi :10.1214/aos/1176344247.

[137] Khan F, Zubek V. Support vector regression for censored data (svrc) : A novel tool for survival analysis. 2008 ; 863–868, doi :10.1109/ICDM.2008.50.

[138] Raykar V, Steck H, Krishnapuram B, Oberije C, Lambin P. On ranking in survival analysis : Bounds on the concordance index. 2007.

[139] Evers L, Messow CM. Sparse kernel methods for high-dimensional survival data. *Bioinformatics (Oxford, England)* 08 2008 ; **24** :1632–8, doi :10.1093/bioinformatics/btn253.

[140] Van Belle V, Huffel S. Support vector machines for survival analysis. *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare* 01 2007 ; .

[141] Van Belle V, Pelckmans K, Huffel S, Suykens J. Support vector methods for survival analysis : A comparison between ranking and regression approaches. *Artificial intelligence in medicine* 08 2011 ; **53** :107–18, doi :10.1016/j.artmed.2011.06.006.

[142] Pölsterl S, Navab N, Katouzian A. Fast training of support vector machines for survival analysis. 2015 ; 243–259, doi :10.1007/978-3-319-23525-7_15.

[143] Lee CP, Lin CJ. Large-scale linear ranksvm. *Neural computation* 01 2014 ; **26**, doi :10.1162/NECO_a_00571.

[144] Pölsterl S, Navab N, Katouzian A. An efficient training algorithm for kernel survival support vector machines 11 2016 ; .

[145] Jamieson K, Talwalkar A. Non-stochastic best arm identification and hyperparameter optimization. *ArXiv and in International Conference on Artificial Intelligence and Statistics AISTATS 2015* 2016 ; **abs/1502.07943**.

[146] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna : A next-generation hyperparameter optimization framework. 2019 ; 2623–2631, doi :10.1145/3292500.3330701.

[147] Wan F. Simulating survival data with predefined censoring rates for proportional hazards models : Simulating censored survival data. *Statistics in Medicine* 11 2016 ; **36**, doi :10.1002/sim.7178.

[148] Tacconelli E. Crd's guidance for undertaking reviews in health care. *Lancet Infectious Diseases - LANCET INFECT DIS* 04 2010 ; **10** :226–226, doi :10.1016/S1473-3099(10)70065-7.

[149] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. *Cochrane Handbook for Systematic Reviews of Interventions* 09 2019 ; doi :10.1002/9781119536604.

[150] NICE. Guide to the methods of technology appraisal. *NICE* 01 2013 ; .

[151] Wu T. Lasso penalized semiparametric regression on high-dimensional recurrent event data via coordinate descent. *Journal of Statistical Computation and Simulation - J STAT COMPUT SIM* 01 2012 ; **83** :1–11, doi :10.1080/00949655.2011.652114.

[152] Zhao H, SUN D, LI G, Sun J. Variable selection for recurrent event data with broken adaptive ridge regression : Variable selection for recurrent event data. *Canadian Journal of Statistics* 08 2018 ; **46**, doi :10.1002/cjs.11459.

[153] Gupta G, Sunder V, Prasad R, Shroff G. *CRESA : A Deep Learning Approach to Competing Risks, Recurrent Event Survival Analysis*. 2019 ; 108–122, doi :10.1007/978-3-030-16145-3_9.

[154] Jing BZ, Zhang T, Wang ZX, Jin Y, Liu K, Qiu WZ, Ke L, Sun Y, He C, Hou D, *et al.*. A deep survival analysis method based on ranking. *Artificial intelligence in medicine* 2019 ; **98** :1–9.

[155] Kim JY, Lee Y, Yu J, Park Y, Lee SK, Lee M, Lee JE, Kim S, Nam S, Park Y, *et al.*. Deep learning-based prediction model for breast cancer recurrence using adjuvant breast cancer cohort in tertiary cancer center registry. *Frontiers in Oncology* 05 2021 ; **11**, doi :10.3389/fonc.2021.596364.

[156] Wang P, Li Y, Reddy CK. Machine learning for survival analysis : A survey. *ACM Computing Surveys* 2019 ; **51 :6** :110.

[157] Bull L, Lunt M, Martin G, Hyrich K, Sergeant J. Harnessing repeated measurements of predictor variables for clinical risk prediction : a review of existing methods. *Diagnostic and Prognostic Research* 07 2020 ; **4**, doi :10.1186/s41512-020-00078-z.

[158] Andersen P, Gill R. Cox's regression model for counting processes : A large sample study. *The Annals of Statistics* 12 1982 ; **10**, doi :10.1214/aos/1176345976.

[159] Prentice R, WILLIAMS B, Peterson A. On the regression analysis of multivariate failure time data. *Biometrika* 08 1981 ; **68**, doi :10.1093/biomet/68.2.373.

[160] Wei L, Lin D, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of The American Statistical Association - J AMER STATIST ASSN* 12 1989 ; **84** :1065–1073, doi :10.1080/01621459.1989.10478873.

[161] Vaupel J, Manton K, Stallard PE. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 08 1979 ; **16** :439–454, doi :10.2307/2061224.

[162] Cox D. Regression models and life-tables (with discussion). *J Royal Statistical Society, Series B* 11 1971 ; **34**, doi :10.1111/j.2517-6161.1972.tb00899.x.

[163] Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* Feb 1997 ; **16**(4) :385–395, doi :10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

[164] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 2011 ; **39**(5), doi :10.18637/jss.v039.i05. URL `http://www.jstatsoft.org/v39/i05/`.

[165] Kawaguchi E, Suchard M, Liu Z, Li G. A surrogate l0 sparse cox's regression with applications to sparse high-dimensional massive sample size time-to-event data. *Statistics in Medicine* 12 2019 ; **39**, doi :10.1002/sim.8438.

[166] Kim S, Schaubel D, McCullough K. A c-index for recurrent event data : Application to hospitalizations among dialysis patients : C-index for recurrent event data. *Biometrics* 08 2017 ; **74**, doi :10.1111/biom.12761.

[167] Jahn-Eimermacher A, Ingel K, Ozga AK, Erdmann S, Binder H. Simulating recurrent event data with hazard functions defined on a total time scale. *BMC Medical Research Methodology* 12 2015 ; **15**, doi :10.1186/s12874-015-0005-2.

[168] Jahn-Eimermacher A. Comparison of the andersen-gill model with poisson and negative binomial regression on recurrent event data. *Computational Statistics and Data Analysis* 07 2008 ; **52** :4989–4997, doi :10.1016/j.csda.2008.04.009.

[169] Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, Liss J. Digital medicine and the curse of dimensionality. *NPJ digital medicine* Oct 2021 ; **4**(1) :153, doi :10.1038/s41746-021-00521-5.

[170] Uno H, Cai T, Pencina M, D'Agostino R, Wei L. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* 05 2011 ; **30** :1105–17, doi :10.1002/sim.4154.

[171] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* 1999 ; **18**(17-18) :2529–2545.

[172] Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* 2006 ; **48**(6) :1029–1040.

[173] Bussy S, Guilloux A, Gaïffas S, Jannot AS. C-mix : A high-dimensional mixture model for censored durations, with applications to genetic data. *Stat Methods Med Res* 2019 ; **28(5)** :1523–1539.

[174] Paioni P, Jäggi V, Tilen R, Seiler M, Baumann P, Bräm D, Jetzer C, Haid R, Goetschi A, Goers R, *et al.*. Gentamicin population pharmacokinetics in pediatric patients—a prospective study with data analysis using the saemix package in r. *Pharmaceutics* 10 2021 ; **13** :1596, doi :10.3390/pharmaceutics13101596.

[175] Brauer F. *Compartmental Models in Epidemiology*, vol. 1945. 2008 ; 19–79, doi :10.1007/978-3-540-78911-6_2.

Chapitre 6

Annexes

6.1 Synthèse du parcours professionnel et contexte d'exercice

Après l'obtention d'un DEA de biomathématiques en 1999 et trois années d'allocataire de recherche entrecoupées par un congé maternité de 2000 à 2004 à l'Université de Paris VI, je fus ATER de 2004 à 2005 au département de mathématiques de l'Université de Nantes. Ma carrière de Maîtresse de Conférences (MCF) a démarré par un poste contractuel de 2005 à 2006 à l'Agrocampus Rennes avant d'obtenir mon poste de MCF à l'Université Rennes 1 en septembre 2006.

J'ai ainsi pu enrichir mon expérience d'enseignement, principalement en Statistique dans les masters de la faculté de médecine, mais aussi dans d'autres départements, université et école d'ingénieurs. Mon expérience en enseignement est ainsi très diversifiée en termes de publics allant des étudiants en 1ère année de médecine, sport ou mathématiques, aux $2^{\text{èmes}}$ années de master, en passant par des licences professionnelles et école d'ingénieurs ENSAI (Ecole Nationale de la Statistique et de l'Analyse de l'Information), avec un service annuel qui dépasse les 192 heures.

Mon implication pédagogique fut très intense dès mon recrutement à la faculté de médecine en termes de responsabilité d'Unités d'Enseignement, mais aussi de **responsabilité de master 1** (M1). En charge officiellement de 2012 à 2016 du M1 Santé Publique j'étais la N+1 de la secrétaire affiliée. J'ai assuré de nombreuses années, la promotion, la coordination importante dans le cadre de co-habilitations avec d'autres universités et l'Ecole des Hautes Etudes en Santé Publique (EHESP), et la co-rédaction de la mention qui était en pleine croissance, en étroite collaboration avec son porteur. Très impliquée dans la création et le développement de cette mention et de l'un de ses masters 2 (Modélisation en Pharmacologie Clinique et Epidémiologie), j'ai très tôt, mesuré l'importance de l'interdisciplinarité, qui s'illustre aujourd'hui par

- un **rattachement à l'UFR Médecine**,
- un **CNU de pharmacie (85, qualifiée en 26 et 67)**,
- une affectation principale dans une **unité mixte du CNRS en mathématiques**,
- et un émargement dans un Centre d'Investigation Clinique (CIC **INSERM**).

Durant les dix premières années de ma carrière, ma recherche s'appuyait sur les besoins des chercheurs du CIC (mon rattachement principal à l'époque), et consistait en l'apport de mes compétences

statistiques dans des domaines aussi variés que la recherche de seuils de biomarqueurs, la méthodologie des essais cliniques, la modélisation linéaire et non linéaire à effets mixtes.

En 2016, grâce à un Congé Recherche et Conversion Thématique, j'ai pu développer de nouvelles collaborations en recherche dans le domaine de la modélisation de propagation d'épidémies (ma thématique de recherche initiale) avec le Laboratoire Traitement du Signal et de l'Image (LTSI - INSERM) et l'Institut de Recherche MAthématique de Rennes (IRMAR - CNRS, que j'ai rejoint en tant que chercheuse associée en 2018, puis en rattachement principal en septembre 2019).

J'ai ainsi co-dirigé la thèse de Canelle Poirier, soutenue en 2019 devant un jury pluridisciplinaire d'experts en statistique, en sciences des données à Santé Publique France, et en informatique médicale dans l'école doctorale Biologie Santé. Elle s'intitulait « Modèles statistiques d'aide à la décision en santé publique basés sur la réutilisation des données massives en santé : application à la surveillance syndromique », et s'attachait à comparer des modèles statistiques et d'apprentissage, pour montrer que la réutilisation des données massives hospitalières pouvait être un réel apport pour la surveillance syndromique. Ce travail s'inscrivait dans le cadre d'un projet ANR, et a bénéficié d'un financement très sélectif pour une mobilité de 6 mois de la doctorante à Harvard.

Cette même année 2019, j'ai obtenu une mobilité (sans décharge d'enseignement) de 6 mois dans l'équipe de statistique de l'Université de Southampton en Angleterre (Southampton Statistical Sciences Research Institute), où j'ai pu continuer à apporter mon expertise en biostatistique avec la publication d'un article sur l'asthme, et initier un travail sur le point de se finaliser sur l'identifiabilité des paramètres de modèles de propagation d'épidémie.

Dès mon retour plusieurs co-encadrements de stages se sont enchaînés sur des comparaisons et évaluations de différentes approches d'apprentissage supervisé sur des données de survie, en collaboration avec l'équipe « Science de l'information au service de la médecine personnalisée » de l'UMR 1138 - Centre de Recherche des Cordeliers (Paris Sorbonne - INSERM). Ainsi, je co-encadre la thèse de Juliette Murris qui démarre, intitulée « Prédictions d'événements récurrents en grande dimension auprès de patients atteints de cancer digestif ». Cette thèse bénéficie d'un financement CIFRE avec Pierre Fabre dans l'école doctorale parisienne Epidémiologie et sciences de l'information biomédicale.

Parallèlement à ma carrière d'enseignante-chercheuse, je suis membre du bureau du Président de Rennes 1 depuis 2008, en tant que :
- **chargée de mission Promotion de la santé et vie sociale** de 2008 à 2012,
- **chargée de mission Qualité de Vie au Travail (QVT)** de 2012 à 2020 et
- **Vice-Présidente (VP) QVT et action sociale** depuis 2020.

Mes lettres de mission successives sont denses (sans décharge d'enseignement pour raison de service à l'UFR médecine) et je participe à de nombreuses commissions pour pouvoir assurer mes fonctions (CHSCT, commission Ressources Humaines, comité d'action sociale, Commission Consultative Paritaire des personnels non-titulaires).

Dès 2008, j'ai accompagné notre Service Inter-universitaire de Médecine Préventive et de Promotion de la Santé qui a obtenu en 2010 l'agrément pour se constituer en centre de santé. Ce fut un dossier important avec la CPAM et de nombreux acteurs. La possibilité de compléter l'activité de prévention par une activité de soins avec prescription permet depuis, une meilleure prise en charge en matière de soins aux étudiants.

J'ai également participé à la création d'un service d'aide à la vie étudiante en 2009, notamment sur la partie « pôle handicap » qui nécessitait une restructuration pour une meilleure visibilité budgétaire et humaine, avec une mise en place compliquée d'une commission plurielle du correspondant handicap, des responsables pédagogiques et de la scolarité, pour décider d'aménagement individuel spécifique assurant l'égalité des chances à tous les étudiants (comme par exemple le tiers temps aux examens).

Un des piliers de notre politique d'amélioration des conditions de travail était la crèche universitaire. Après une évaluation des besoins, de nombreuses pistes ont été envisagées (la création d'une crèche avec recherche de partenaires, le partenariat public/privé pour une construction, la gestion associative plutôt que privée pour un montage éventuel avec la ville de Rennes, la réservation de places en offre réseau pour offrir ce service également aux sites délocalisés), avant que je puisse négocier la réservation de quelques places à un coût intéressant conventionné avec la CAF. Depuis 2011, de nombreuses familles ont pu bénéficier de places en crèche. Onze ans plus tard, je m'apprête à conduire une nouvelle évaluation des besoins dans le cadre de notre politique de site rennais.

Entre 2010 et 2012, j'ai été **élue membre du Conseil d'Administration et de la commission des finances**, en plein passage effectif à l'autonomie des universités.

En 2012, j'ai restreint ma mission au personnel. Trois actions ont rapidement vu le jour : la mise en place d'un parrainage par un senior pour les nouveaux MCF, la rédaction d'un livret d'accueil sur le fonctionnement de l'Université, et la constitution d'un groupe de travail QVT. Je coordonne ce groupe depuis bientôt 10 ans, au départ avec le conseiller de prévention, et aujourd'hui avec la responsable du pôle QVT, dialogue social et action sociale. Ce groupe est à géométrie variable : parfois intra-universitaire, parfois je l'élargis à d'autres établissements universitaires, écoles et organismes, selon les sujets. Des représentants du CHSCT en font partie, nous avons construit ensemble de nombreux projets comme une formation management obligatoire pour les nouveaux responsables d'équipe avec une forte sensibilisation aux risques psychosociaux (RPS), une enquête sur la pause méridienne des personnels en collaboration avec le CROUS, et surtout l'élaboration de notre plan d'actions QVT au travers d'une méthodologie participative. Ce dernier a démarré en 2018 par un questionnaire, des séminaires annuels théâtraux (le premier sur les RPS et le 2ème sur le télétravail subi ou choisi), des ateliers ouverts à tous, un questionnaire de fin de 1er confinement, et un budget participatif. Une réflexion est en cours pour modifier le fonctionnement avec un groupe Rennes 1 et un groupe avec mes homologues des établissements du site.

J'ai bien sûr traité des dossiers en dehors de ce groupe tels que la mise en place d'une cellule de veille au travail (assistante sociale, médecin de prévention et DRH qui travaillent sur les situations individuelles), une convention et des projets avec le réseau PAS (Prévention, Aide, Suivi) de la MGEN, une participation à la rédaction du plan d'actions sur l'égalité professionnelle entre les femmes et les hommes, et le pilotage du schéma directeur handicap. Je fais également partie de l'équipe d'accueil (interne à Rennes 1) en charge des signalements de harcèlement, discriminations, Violences Sexistes et Sexuelles (VSS).

En tant que VP je travaille aussi avec la direction du Cabinet et le VP en charge des Ressources Humaines et du dialogue social, à l'élaboration du projet d'établissement et du contrat quinquennal.

Mon implication administrative « locale » n'est pas en reste. Pour l'UFR Médecine : j'ai été **élue membre du Comité Pédagogique** (2011-2013), je suis depuis 2008 interlocutrice pédagogique de la Direction du Système d'Information, et depuis 2020 représentante sur les questions de développement durable. Pour l'IRMAR : je suis membre de la commission parité depuis 2020, et membre d'un groupe de travail bimensuel composé des directions d'enseignement et recherche de mathématiques et des collègues

présents dans les conseils de Rennes 1.

Pour assurer mes missions, j'ai eu à cœur de me former sur le handicap, les RPS et les VSS (conférences nationales de l'AMUE, MESRI, ESEN et formations rennaises), mais aussi sur la pédagogie en classe virtuelle et en anglais. J'ai également développé un module en ligne de maîtrise de la rédaction scientifique d'un rapport de projets de master aidée par une ingénieure pédagogique.

6.2 Formation initiale

Année	Diplôme	Etablissement
2004	Doctorat de Biomathématiques (qualification CNU 26 et 67)	Univ. Paris VI
2003	DU Formation Supérieure Biomédicale (IFSBM)	Univ. Paris XI
2000	DEA de Biomathématiques	Univ. Paris VI
1999	Maîtrise Mathématiques Appliquées aux Sciences Sociales	Univ. Paris V
1999	MST Info. et Stat. Appliquées aux Sciences de l'Homme	Univ. Paris V
1997	DUT Statistique et Traitement Informatique des Données	IUT, Vannes

6.3 Activité scientifique complète

6.3.1 Présentation synthétique de l'évolution des thématiques de recherche

J'ai soutenu fin 2004 ma thèse intitulée "Modélisation et analyse de la co-circulation de virus grippaux : diffusion en population, variabilité génomique et impact clinique". Le tableau ci-dessous liste et situe temporellement mes thématiques de recherche à partir de mes publications (en noir) et projets en cours (en gris) (tableau 6.1).

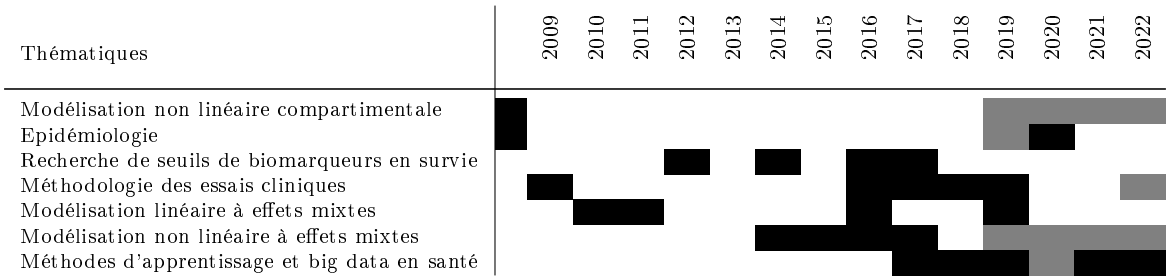


Table 6.1 – Evolution des thèmes de recherche

Les thématiques de ma thèse (avec la bioinformatique que j'ai mis de côté depuis, mais qui avait nécessité une formation spécifique) étaient la modélisation compartimentale que j'enseigne toujours à l'ENSAI et reprends depuis 2019 avec un projet que je porte en collaboration avec deux chercheurs anglais ; et l'épidémiologie naturellement. Sur ce dernier thème, j'ai d'ailleurs coécrit récemment un article sur l'asthme en collaboration internationale [2].

Recherche de seuils de biomarqueurs en survie

A mon arrivée en tant que MCF à Rennes 1, j'étais rattachée à l'Unité Fonctionnelle de Biométrie du CIC, j'étais alors sollicitée pour l'analyse statistique sortant du champ de compétences des biostatisticiens de l'unité, comme les modèles de survie discriminant en groupes à faible risque et à haut risque par identification de seuils sur une variable continue. J'ai ainsi travaillé en cancérologie avec le Centre Eugène Marquis. Ces travaux ont été décrits au chapitre 3, paragraphe 3.2.

Méthodologie des essais

Le CIC permet aux médecins du CHU de Rennes de réaliser leurs projets de recherche clinique, de la conception du protocole (méthodologie) à sa valorisation (publication scientifique), en incluant la réalisation des investigations cliniques et l'analyse des données. Travailler dans cet univers m'a aussi donné l'opportunité d'enseigner 10 ans la méthodologie et la statistique pour la recherche clinique en licence professionnelle Statistique et Informatique pour la Santé à Vannes. Certains essais peuvent être expérimentaux sur animaux. J'ai participé à une étude sur 105 cobayes qui ont été soumis à une sténose de l'aorte thoracique au moyen d'aiguilles de 2 calibres ou d'opérations fictives similaires (4 groupes). L'article publié décrit les caractéristiques des ventricules gauches et droits et leur évolution dans le temps [17]. Cependant, ma plus grande contribution en tant que méthodologiste d'essais thérapeutiques est la couverture de 8 congrès européens et américains d'hématologie en binôme clinicien-statisticien, le but étant d'éclairer en direct le clinicien sur la méthodologie et les résultats de topos choisis, puis de rédiger conjointement les résumés et critiques méthodologiques, enfin il s'agit aussi de vulgariser des méthodes nouvelles ou complexes dans des focus statistiques (Numéros spéciaux Regards croisés cliniciens statisticiens d'Horizons Hémato [25].

Modélisation linéaire à effets mixtes

Rapidement mon soutien à l'Unité Fonctionnelle de Biométrie du CIC s'est focalisé sur les modèles non linéaires à effets mixtes. Dans le cadre de mesures répétées dans le temps, d'essais cliniques en cross-over, ou de méta-analyses, il est souvent nécessaire de rajouter des effets aléatoires, en plus des effets fixes dans les modèles linéaires pour tenir compte des corrélations engendrées. J'ai ainsi participé à l'analyse de 4 études et publications sur des sujets variés. Deux études portaient sur de la pharmacologie : une ayant permis d'avancer sur l'évaluation des effets hémodynamiques de faibles doses d'hydrocortisone et de fludrocortisone sur le pronostic après choc septique [16], et l'autre ayant montré que les patients transplantés hépatiques présentant une récidive de l'hépatite C qui initient de la ribavirine associée à un régime antiviral à action directe sofosbuvir-daclatasvir peuvent être à risque de concentrations de tacrolimus plus faibles en raison d'une anémie probable induite par la ribavirine et d'un score de fibrose plus élevé [3]. La troisième analyse que j'ai réalisée avec un modèle mixte a permis d'estimer la fréquence des pneumonies interstitielles chez les patients atteints de carcinome hépatocellulaire recevant de l'iode-131 Lipiodol. A partir d'une méta-analyse de 36 études, nous avons montré qu'elle semble plus élevée et plus précise que celle estimée précédemment. Le risque semble être lié au nombre d'injections et au niveau de dose par injection [15]. Cet article publié m'a conduit quelques années plus tard à me charger d'un cours de méta-analyse en master 2, que j'assure encore aujourd'hui. Le dernier article ayant nécessité mes compétences en modèle linéaire à effets mixtes est issu d'une collaboration entre l'ENS Rennes et l'équipe "Ischémie, macro et microcirculation" du CIC. Pour débroussailler le terrain, j'ai encadré 3 stagiaires de master 1 durant un mois pour étudier les différentes mesures de consommation énergétique à partir d'un échantillon de 30 adultes jeunes et en bonne santé randomisés en cross-over sur des marches à différentes vitesses sur des terrains de pentes différentes. La précision des prédictions utilisant des données de vitesse et de pente de référence (celles réelles connues sur ce terrain, non GPS) était élevée. La précision à partir des données GPS pentes non corrigées diminuait même si elle restait substantielle. La précision a été

grandement améliorée lorsque les données corrigées avec le logiciel de projection cartographique ont été utilisées. Ces résultats offrent des perspectives prometteuses et applications cliniques liées à l'évaluation de la dépense énergétique pendant la marche libre [8].

Modélisation non linéaire à effets mixtes

Avec la restructuration du CIC et la création d'une équipe Modélisation Pharmacocinénique (PK) et Pharmacocinénique-Pharmacodynamique (PK-PD) en 2012, je me suis intéressée aux modèles non linéaires à effets mixtes. La première étude sur laquelle j'ai travaillé fait partie de mes publications les plus significatives sur la modélisation de la croissance tumorale [12]. J'ai utilisé l'algorithme SAEM (Stochastic Approximation Expectation-Maximisation) qui maximise la vraisemblance et converge avec succès pour des modèles complexes. Cet algorithme est rapide, peu sensible à l'initialisation, et très utilisé pour la modélisation PK et PK-PD. Il était alors disponible uniquement sur le logiciel Monolix. Nous avons décidé de l'implémenter sous R avec les chercheurs ayant développé SAEM. Le package saemix est disponible sur Rcran et a fait l'objet d'une publication [6]. Ceci a été expliqué au chapitre 3 paragraphe 3.2.1.
J'ai ensuite participé à l'encadrement d'un stage de master 2 Ingénierie mathématiques option statistique de l'Université de Nantes sur l'étude HEPMEN, qui s'est poursuivit par un contrat d'ingénieur de 6 mois, pour étudier la dynamique conjointe de l'hepcidine et du fer sérique pendant le cycle menstruel [10,11]. Le modèle conjoint non linéaire à effets mixtes incluant un modèle structural compartimental est présenté dans la partie 3.2.2.

Méthodes d'apprentissage et big data en santé

A l'occasion du co-encadrement de la thèse de Canelle Poirier dans la continuité de son stage double cursus de masters, nous avons travaillé sur un sujet liant modélisation et big data en santé sur l'Entrepôt de données biomédicales de l'HOPital de Rennes (eHOP). C'est une base de données massives qui centralise l'exhaustivité des informations des patients du CHU de Rennes. Ce sujet m'a permis de revenir à ma thématique de recherche initiale sur la modélisation d'épidémies, et m'a fait évoluer vers les méthodes d'apprentissage de par les données massives intégrées. La thèse fut soutenue en 2019 devant un jury pluridisciplinaire d'experts en statistique, en sciences des données à Santé Publique France, et en informatique médicale. Ce travail s'inscrivait dans le cadre du projet ANR INSHARE (2016-2019 INtegrating and SHAring health data for REsearch), et a intégré une mobilité de 6 mois de la doctorante à Harvard bénéficiant du financement très sélectif Fulbright U.S. Student Program. Trois articles sont aujourd'hui publiés, et un en révision [1,4,5,27], les modèles et résultats sont détaillés dans le chapitre 2. Cette thèse a démarré en même temps que mon CRCT, me permettant une implication complète et le développement de mes propres projets, notamment en préparant ma mobilité anglaise de 6 mois (en 2019) avec un projet plus fondamental en statistique que nous finalisons actuellement. Partant d'applications grippe et gastro-entérite pour la thèse, je m'intéresse actuellement à la COVID19. Ceci est décrit dans la partie 3.2.3. J'ai en parallèle pris une nouvelle direction suite aux compétences acquises lors de cette thèse où trois modèles statistiques et de machine learning ont été comparés, en montant une collaboration avec l'équipe « Science de l'information au service de la médecine personnalisée » de l'UMR 1138 - Centre de Recherche des Cordeliers (Paris Sorbonne - INSERM). Nous avons démarré par des co-encadrements de stages de master 2 sur des comparaisons de méthodes de machine learning en survie. Nous travaillons actuellement sur des simulations pour finaliser un premier papier dont la première partie a été présentée dans deux congrès et publiée dans des actes [23]. En parallèle, nous avons obtenu une bourse CIFRE pour financer la thèse de Juliette Murris sur des prédictions d'événements récurrents en grande dimension auprès de patients atteints de cancer digestif. Un premier papier est déjà accepté sous réserve de révisions mineures [28]. Ceci a été décrit dans les parties 4.2 et 4.3 du chapitre 4.

6.3.2 Publications et productions scientifiques : une trentaine de travaux dont les 16 décrits pour l'HDR

Au total, j'ai coécrit 24 articles publiés ou acceptés récemment (dont 19 depuis que je suis MCF [1-17,27,28]) dans des revues internationales à comité de lecture, ainsi que 3 actes publiés de congrès [23-25] dont un résumé long de 6 pages, 8 numéros de la revue française des pratiques en hématologie (Regards croisés cliniciens statisticiens d'Horizons Hémato [26]), 1 article soumis dans une revue internationale à comité de lecture [29], et 1 présentation en congrès international d'un article en fin d'écriture [32].

Articles publiés dans des revues internationales à comité de lecture

Les auteures soulignées sont les deux doctorantes.

[1] Poirier C, Hswen Y, Bouzillé G, Cuggia M, **Lavenu A**, Brownstein J. S, Brewer T, Santillana M. Influenza forecasting for French regions combining EHR, web and climatic data sources with a machine learning ensemble approach. **PLoS One. 2021** May 19 ;16(5) :e0250890. doi : 10.1371/journal.pone.0250890.

[2] Azim A, Freeman A, **Lavenu A**, Mistry H, Haitchi H. M, Newell C, Cheng Y, Thirlwall Y, Harvey M, Barber C, Pontoppidan K, Dennison P, Arshad H, Djukanovic R, Howarth P, Kurukulaaratchy R. J. New Perspectives on Difficult Asthma ; Sex and Age of Asthma-Onset Based Phenotypes. **J.A.C.I. 2020** Jun 13.

[3] Barrail-Tran A, Goldwirt L, Gelé T, Laforest C, **Lavenu A**, Danjou H, Radenne S, Leroy V, Housset-Debry P, Duvoux C, Kamar N, De Ledinghen V, Canva V, Conti F, Durand F, D'Alteroche L, Botta-Fridlund D, Moreno C, Cagnot C, Samuel D, Fougerou-Leurent C, Pageaux GP, Duclos-Vallée JC, Taburet AM, Coilly A. Comparison of the effect of direct-acting antiviral with and without ribavirin on cyclosporine and tacrolimus clearance values : results from the ANRS CO23 CUPILT cohort. **Eur J Clin Pharmacol. 2019** Nov. 75(11) :1555-1563. doi : 10.1007/s00228-019-02725-x.

[4] Poirier C, **Lavenu A**, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods : Comparison Study. **J Med Internet Res. JMIR Public Health Surveill. 2018** Dec 21 ;4(4) :e11361. doi : 10.2196/11361.

[5] Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, **Lavenu A**, Cuggia M. Leveraging hospital big data to monitor flu epidemics. **Computer Methods and Programs in Biomedicine. 2017** Nov 15.

[6] Comets E, **Lavenu A**, Lavielle M. Parameter Estimation in Nonlinear Mixed Effect Models Using saemix, an R Implementation of the SAEM Algorithm. **Journal of Statistical Software. 2017** Aug 29. doi : 10.18637/jss.v080.i03.

[7] Quillien V, **Lavenu A**, Ducray F, Meyronet D, Chinot O, Fina F, Sanson M, Carpentier C, Karayan-Tapon L, Rivet P Entz-Werle N, Legrain M, Lechapt-Zalcman E, Levallet G, Escande F, Ramirez C, Chiforeanu D, Vauleon E and Figarella-Branger D. Clinical validation of the ce-ivd marked therascreen mgmt kit in a cohort of glioblastoma patients. **Cancer Biomarkers. 2017** Aug 3. vol. doi : 10.3233/CBM-170191. Preprint, no. Preprint, pp. 1-7.

[8] De Müllenheim PY, Dumond R, Gernigon M, Mahé G, **Lavenu A**, Bickert S, Prioux J, Noury-Desvaux B, Le Faucheur A. Predicting metabolic rate during level and uphill outdoor walking using a low-cost GPS receiver. **J Appl Physiol (1985). 2016** Aug 1 ;121(2) :577-88. doi : 10.1152/japplphysiol.00224.2016. Epub 2016 Jul 8.

[9] Quillien V, **Lavenu A**, Ducray F, Joly MO, Chinot O, Fina F, Sanson M, Carpentier C, Karayan-Tapon L, Rivet P, Entz-Werle N, Legrain M, Lechapt Zalcman E, Levallet G, Escande F, Ramirez C, Chiforeanu D, Vauleon E, and Figarella-Branger D. Validation of the high-performance of Pyrosequencing for clinical MGMT testing on a cohort of glioblastoma patients from a prospective dedicated multicentric trial. **Oncotarget. 2016** Sep 20 ;7(38) :61916-61929.

[10] Angeli A, Laine F, **Lavenu A**, Ropert M, Lacut K, Gissot V, Sacher-Huvelin S, Jezequel C, Moignet A, Laviolle B, Comets E. Joint model of iron and hepcidin during the menstrual cycle in healthy women. **AAPS J. 2016** Mar ;18(2) :490-504. doi : 10.1208/s12248-016-9875-4. Epub 2016 Feb 2.
[11] Lainé F, Angeli A, Ropert M, Jezequel C, Bardou-Jacquet E, Deugnier Y, Gissot V, Lacut K, Sacherhuvelin S, **Lavenu A**, Laviolle B, Comets E. Variations of hepcidin and iron-status parameters during menstrual cycle in healthy women. **Br J Haematol. 2015** Dec 21. doi : 10.1111/bjh.13906. [Epub ahead of print] No abstract available.
[12] Verhoest G, Dolley-Hitze T, Jouan F, Belaud-Rotureau MA, Oger E, **Lavenu A**, Bensalah K, Arlot-Bonnemains Y, Collet N, Rioux-Leclercq N, Vigneau C. Sunitinib combined with angiotensin-2 type-1 receptor antagonists induces more necrosis : a murine xenograft model of renal cell carcinoma. **Biomed Res Int. 2014** ;2014 :901371. doi : 10.1155/2014/901371. Epub 2014 May 22.
[13] Quillien V, **Lavenu A**, Sanson M, Legrain M, Dubus P, Karayan-Tapon L, Mosser J, Ichimura K, Figarella-Branger D. Outcome-based determination of optimal pyrosequencing assay for MGMT methylation detection in glioblastoma patients. **J Neurooncol. 2014** Feb ;116(3) :487-96. doi : 10.1007/s11060-013-1332-y. Epub 2014 Jan 14.
[14] Quillien V, **Lavenu A**, Karayan-Tapon L, Carpentier C, Labussière M, Lesimple T, Chinot O, Wager M, Honnorat J, Saikali S, Fina F, Sanson M, Figarella-Branger D. Comparative assessment of 5 methods (methylation-specific polymerase chain reaction, MethyLight, pyrosequencing, methylation-sensitive high-resolution melting, and immunohistochemistry) to analyze O6-methylguanine-DNA-methyltranferase in a series of 100 glioblastoma patients. **Cancer. 2012** Sep 1 ;118(17) :4201-11. doi : 10.1002/cncr.27392. Epub 2012 Jan 31.
[15] Oger E, **Lavenu A**, Bellissant E, Garin E, Polard E. Meta-analysis of interstitial pneumonia in studies evaluating iodine-131-labeled lipiodol for hepatocellular carcinoma using exact likelihood approach. **Pharmacoepidemiol Drug Saf. 2011** Sep ;20(9) :956-63.
[16] Laviolle B, Le Maguet P, Verdier MC, Massart C, Donal E, Lainé F, **Lavenu A**, Pape D, Bellissant E. Biological and hemodynamic effects of low doses of fludrocortisone and hydrocortisone, alone or in combination, in healthy volunteers with hypoaldosteronism. **Clin Pharmacol Ther. 2010** Aug ;88(2) :183-90.
[17] Laviolle B, Pape D, Verdier MC, **Lavenu A**, Bellissant E. Hemodynamic and histomorphometric characteristics of heart failure induced by aortic stenosis in the guinea pig : comparison of two constriction sizes. **Can J Physiol Pharmacol. 2009** Nov ;87(11) :908-14.
[18] Carrat F, **Lavenu A**. Heterosubtypic immunity to influenza : right hypothesis, wrong comparison. **J Infect Dis. 2006** Jun 1 ;193(11) :1613 ; author reply 1613-4.
[19] Carrat F, **Lavenu A**, Cauchemez S, Deleger S. Repeated influenza vaccination of healthy children and adults : borrow now, pay later ? **Epidemiol Infect. 2006** Feb ;134(1) :63-70.
[20] **Lavenu A**, Leruez-Ville M, Chaix ML, Boelle PY, Rogez S, Freymuth F, Hay A, Rouzioux C, Carrat F. Detailed analysis of the genetic evolution of influenza virus during the course of an epidemic. **Epidemiol Infect. 2006** Jun ;134(3) :514-20. Epub 2005 Nov 29.
[21] Viboud C, Boëlle PY, Cauchemez S, **Lavenu A**, Valleron AJ, Flahault A, Carrat F. Risk factors of influenza transmission in households. **Br J Gen Pract. 2004** Sep ;54(506) :684-9.
[22] **Lavenu A**, Valleron AJ, Carrat F. Exploring cross-protection between influenza strains by an epidemiological model. **Virus Res. 2004** Jul ;103(1-2) :101-5.

Actes publiés de congrès

[23] **Lavenu A**, Murris J, Mareau A, Rouzé T, Fromont M, Gares V, Katsahian S. Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation. **53ème Journées de Statistique de la SFdS. 2022** Juin ; p879-885. 13-17 juin 2022, Lyon, France.

[24] Laviolle B, **Lavenu A**, Kerangueven A-C, Comets E, Bellissant E. Population pharmacodynamic analysis of effect of hydrocortisone and fludrocortisone on phenylephrine-mean arterial pressure dose-response relationship in healthy volunteers. **Fundamental and Clinical Pharmacology. 2012** April ; 26 :26-27. 7ème Congrès de Physiologie, de Pharmacologie et de Thérapeutique (P2T), 4-6 avril 2012, Dijon, France.
[25] Fougerou C, **Lavenu A**, Verdier MC, Tribut O, Bellissant E. Analysis of mycophenolic acid pharmacokinetics at 3 weeks and 3 months after liver transplantation by non-compartmental and individual modelling approaches. **Fundam Clin Pharmacol. 2009** ; 23(S1) : 21. Congrès de Physiologie, de Pharmacologie et de Thérapeutique (P2T), 15-17 avril 2009, Marseille, France.

Articles dans des revues françaises

[26] Numéros spéciaux qui ont couvert l'EHA (European Hematology Association Congress) de Copenhague en 2016, Madrid en 2017, de Stockholm en 2018, d'Amsterdam en 2019, et d'édition virtuelle en 2020 ; et l'ASH (American Society of Hematology) d'Atlanta en 2017, d'Orlando en 2019, et d'édition virtuelle en 2020 : Regards croisés cliniciens statisticiens d'Horizons Hémato – la revue des pratiques en hématologie.

Articles acceptés récemment, ou acceptés avec révisions mineures, ou soumis dans des revues internationales à comité de lecture

[27] C. Poirier, G. Bouzillé, V. Bertaud, M. Cuggia, M. Santillana, **A. Lavenu**. Gastroenteritis forecasting assessing the use of web and EHR data with machine learning approaches. **(accepté fin 2022 à JMIR Public Health and Surveillance)**
[28] J. Murris, A. Charles-Nelson, **A. Lavenu**, S. Katsahian. Towards Filling the Gaps around Recurrent Events in High Dimensional Framework : Literature Review and Early Comparison. **(accepté avec révisions mineures début janvier 2023 dans Biostatistics and Epidemiology)**
[29] C. Nevoret, Y. Tran, S. Guendouz, **A. Lavenu**, T. Damy, S. Katsahian, AI. Tropeano. Healthcare trajectories and mortality in heart failure. **(soumis à European journal of heart failure)**

Présentation dans des congrès

[30] ***Lavenu A***, Murris J, Mareau A, Rouzé T, Fromont M, Gares V, Katsahian S. Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation. 53ème Journées de Statistique de la SFdS. 13-17 juin 2022, Lyon, France.
[31] ***Lavenu A***, Murris J, Mareau A, Rouzé T, Fromont M, Gares V, Katsahian S. Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation. Congrès Intelligence Artificielle et Santé : approches interdisciplinaires. 29 juin - 1 juillet 2022, Nantes, France.
[32] *Lemaitre F*, **Lavenu A**, Coste G, Tron C, Ferrand-Sorre MJ, Lalanne S, Franck B, Vigneau C, Verdier MC, Bellissant E, Chemouny J. Does inflammation influence tacrolimus pharmacokinetics in kidney transplantation ? 19ème congrès international IATDMCT (International Association of Therapeutic Drug Monitoring and Clinical Toxicology). 19-22 septembre 2021, Rome, Italie. (présentation orale*)
[33] ***Lavenu A***, Comets E et Lavielle M. Saemix, une version R de l'algorithme SAEM pour l'estimation de paramètres dans les modèles non linéaires à effets mixtes. 1ères Rencontres R Juillet 2012 Bordeaux. (présentation orale*)

[34] ***Lavenu A*** et Causeur D. Estimation d'un modèle allométrique en présence d'observations manquantes de la covariable. XXXVIIèmes Journées de la Société Française de Statistique mai 2006 Clamart. (présentation orale*)
[35] ***Lavenu A***, Leruez-Ville M, Chaix ML, Boelle PY, Rogez S, Freymuth F, Hay A, Rouzioux C, Carrat F. Detailed analysis of the genetic evolution of influenza virus during the course of an epidemic. The International Conference on Options for the Control of Influenza V Octobre 2003 Okinawa (Japon). (présentation poster*)
[36] ***Lavenu A***, Valleron AJ, Carrat F. Exploring cross-protection between influenza strains by an epidemiological model. The first European Influenza Conference Octobre 2002 Malte. (présentation orale*)

* présentation par l'auteur en italique

6.3.3 Encadrement doctoral et scientifique

Thèse soutenue

Canelle Poirier : Thèse du 1er septembre 2016 au 13 juin 2019. Co-encadrement à 50 % avec Pr Valérie Bertaud (HDR) et Dr Guillaume Bouzillé (AHU doctorant). 4 articles ont été publiés et décrits au chapitre 2 paragraphes 2.2 et 2.3.
Depuis juin 2021 : Canelle est Prestataire INSEP (Laboratoire SEP), elle travaille sur l'élaboration d'un cahier des charges pour la création d'un système de gestion, de visualisation et de partage des données, et en post-doctorat à Rennes 2 avec Magalie Fromont. Après un premier post-doctorat à Harvard Medical School - Boston Children's Hospital (juillet 2019 à août 2021), où elle a travaillé sur la modélisation de différentes épidémies telles que la Dengue, la grippe et la COVID-19 en réutilisant des sources de données massives (Web, hospitalières, climatiques) à l'aide de methodes de machine learning, elle a pu travailler sur l'évaluation des mesures contre la COVID-19 en France (Post-Doctorat INSERM - Epicx lab, avril 2021 - janvier 2022).

Thèse en cours

Juliette Murris : Début de thèse au 1er septembre 2021, en co-encadrement à 50%, avec Pr Sandrine Katsahian. Un article est déjà accepté sous réserve de révisions mineures et un article en fin d'écriture suite au travail de master 2 et du co-encadrement d'un stagiaire de master 1. Le financement est une bourse CIFRE avec l'entreprise Pierre Fabre qui la fait travailler sur des études en parallèle de la thèse.

Encadrement ou co-encadrement de stages

Stages de master 2 Santé Publique, parcours Modélisation en pharmacologie Clinique et Epidémiologie (MPCE), Université de Rennes 1 : 8 étudiantes
- Cécile Crolard (2019) Comparaison de méthodes pour données de survie de grande dimension.
- Jing Wang (2013) Modélisation PK du mycophénolate mofétil après transplantation hépatique.
- Zinnya Del Villar (2012) Efficacité de la kiné respiratoire contre la bronchiolite aigüe.
- Anne-Cécile Kerangueven (2011) Modélisation, par approche de population, de la réponse pressive à la phényléphrine, chez le volontaire sain en condition d'hypoaldostéronisme.
- Pascale Le Maguet (2010) Effets de faibles doses de fludrocortisone et d'hemisuccinate d'hydrocortisone, administrées seules ou associées, sur la réponse pressive à la phenylephrine, dans des conditions de

suppression de l'aldosterone endogene, chez le volontaire sain.
- Fabienne Le Gac (2009) Méthodologie des essais de phase I en cancérologie : Simulation par évènements en temps discrets et proposition d'un nouveau schéma expérimental.
- Claire Fougerou (2008) Modélisation de la pharmacocinétique du mycophénolate mofénil 3 semaines et 3 mois après une transplantation hépatique.
- Chiraz Hamila (2007) Évaluation de facteurs modulant l'intervalle entre les transfusions chez des patients ayant une leucémie aiguë ou une autogreffe.

Autres stages de master 2 de statistique ou données massives en santé : 6 étudiants
- Juliette Murris (2021, *Données Massives en Santé, Université de Paris*) Événements récurrents en grande dimension : état de l'art et comparaison de méthodes.
- Anne-Isabelle Tropeano (2020, *Données Massives en Santé, Université de Paris*) Trajectoires de soins pronostiques du décès chez des patients insuffisants cardiaques.
- Claire Daboudet (2020, *Maths Appli, Stat (MAS), parcours Science des données (SDD), UR2*) Étude des flux aux urgences de Brest.
- Alexis Mareau (2020, *Maths Appli, Stat (MAS), parcours Science des données (SDD), UR2*) Comparaison de méthodes pour données de survie de grande dimension.
- Canelle Poirier (2016, *double cursus Stat Appli pour l'entreprise UR2 et MTIBH UR1*) Les données massives hospitalières pour la surveillance des épidémies de grippe.
- Adeline Angeli (2014, *Ingénierie mathématique, parcours Statistique, Université de Nantes*) Modélisation des variations sériques d'hepcidine chez la femme non ménopausée.

Stage de master 2 Psychologie du travail et des organisations, UR2 : 1 étudiante
- Sabine Dessaint (2018) L'accompagnement du plan de modernisation et de développement de l'Université de Rennes 1.

Stages en cours de cursus de 1 à 3 mois : 9 étudiants
- Timothé Rouzé (2021, *M1 Santé Publique, UR1*) L'optimisation des hyperparamètres pour les méthodes de machine learning.
- Hiba Hamdi (2018, *2ème année ENSAI*) Analyse d'un questionnaire sur la Qualité de Vie au Travail (QVT) à l'Université de Rennes 1.
- Loic Lemariey (2018, *2ème année ENSAI*) Analyse d'un questionnaire sur la Qualité de Vie au Travail (QVT) à l'Ecole Normale Supérieure (ENS).
- Ambre Fouché (2017, *L3 MIASHS, UR1*) Évaluation des bonnes pratiques des modèles linéaires mixtes.
- Elie Chedemail (2016, *1ère année de Magistère Statistique et Modélisation Économique, UR1*) Évaluation de modèles linéaires mixtes : AIC, R2, leave-one-out.
- Canelle Poirier + Marie-Laure Aubert + Mélanie Chabot (2015, *M1 Stat Appli pour l'entreprise, UR2*) Modélisation de la dépense énergétique en course à différentes vitesses et pentes.
- Diane Chadaigne (2011, *2ème année ENSAI*) Comparaison de modèles de survie dans une étude liée aux risques de rejet du greffon.

6.3.4 Diffusion et rayonnement

Expertise

L'IRT BCom a lancé une étude de faisabilité de prédiction des entrées aux urgences. De par mon expérience en modèle explicatif des épidémies, en modèle prédictif de séries temporelles, et en machine learning, j'ai accompagné cette équipe sur l'année universitaire 2019-2020 sous un contrat de mise à

disposition pris sur mon temps de recherche. L'idée était de modéliser le nombre de patients se présentant en service d'Urgences que l'on sait de nature purement cyclique (période de l'année, jour de la semaine, heure) mais aussi de nature plus exceptionnelle ou à temporalité variable (épidémies de grippe, de gastro-entérite, conditions météorologiques, pics de pollution, évènement local, etc). Je leur ai également trouvé une stagiaire de master 2 que j'ai co-encadrée. Cette année leur a permis de déposer un projet de plus grande envergure.

Participation jurys de thèse (hors établissement) : en tant qu'examinatrice

- Anaïs Charles-Nelson (24 juin 2020, école doctorale n°393 Santé publique : épidémiologie et sciences de l'information biomédicale, Université de Paris) Traitements des réhospitalisations par des méthodes statistiques prenant en compte les événements récurrents.

- Marie De Antonio (23 juin 2020, école doctorale n°393 Santé publique : épidémiologie et sciences de l'information biomédicale, Université de Paris) Statistiques et modèles de survie pour améliorer la connaissance d'une maladie rare, la dystrophie myotonique.

Diffusion du savoir (vulgarisation)

En parallèle, j'ai participé à des opérations de diffusion du savoir avec une explication des tests d'interaction et conditions de validité des tests en symposium de congrès d'hématologues (Liffré 2017, Monaco 2018, Paris 2018), et la partie « focus statistique » des 8 parutions de regards croisés que j'ai co-écrits pour ce public d'hématologues [25].

Organisation colloques, conférences, journées d'étude

- En 2018-2019, j'ai été membre du comité d'organisation des Journées de la Statistique Rennaises (JSTAR), la 15ème édition de 2019 était sur le thème « Statistique et données de santé ».

- En 2021-2022, je suis membre du comité d'organisation du Colloque Francophone International sur l'Enseignement de la Statistique (CFIES). L'édition de 2022 aura lieu à Rennes et est organisé par la Société Française de Statistique (SFdS).

Invitations dans des universités étrangères

De mars à septembre 2019, j'étais invitée à l'Université de Southampton dans le laboratoire S3RI (Southampton Statistical Sciences Research Institute). J'avais, cette année-là, réalisé mes 192 HETD avant de partir, et j'ai pu sur place me consacrer pleinement à la recherche dans l'équipe anglaise. Je leur ai apporté un projet sur l'identifiabilité des paramètres de modèles de propagation d'épidémie que je continue de porter avec deux chercheurs anglais. J'ai aussi travaillé là-bas avec deux chercheurs cliniciens en publiant l'article sur l'asthme précédemment cité [2].

6.4 Responsabilités collectives et d'intérêt général

6.4.1 Responsabilités administratives

Depuis 14 ans je suis membre du bureau du Président de Rennes 1 (avec réunion a minima mensuelle du bureau entier), alors nommée chargée de mission Promotion de la santé et vie sociale, puis 4 ans plus tard chargée de mission Qualité de Vie au Travail (QVT), finalement pour 2 mandats de 4 ans. En 2020, pour légitimer et fluidifier mon travail avec les services de l'Université, j'ai été nommée vice-présidente Qualité de Vie au Travail et action sociale, déléguée auprès du Vice-Président en charge des Ressources Humaines et du dialogue social. Je participe à de nombreuses commissions pour pouvoir assurer mes fonctions (CHSCT, commission Ressources Humaines, comité d'action sociale, Commission Consultative Paritaire des personnels non-titulaires), et j'ai eu en charge de nombreux dossiers comme la création du centre de santé, la cellule handicap, la crèche universitaire, le groupe QVT, le parrainage des MCF, la formation management, la convention PAS avec la MGEN,...

6.4.2 Responsabilités et mandats locaux

Participation aux conseils centraux

Entre 2010 et 2012, j'ai été élue membre du Conseil d'Administration et de la commission des finances de l'Université. Cette responsabilité de vote de toutes les décisions importantes de l'Université s'est présentée dans un contexte de profonde mutation, marqué notamment par la mise en place de la loi LRU et le passage à l'autonomie effective au premier janvier 2010.
Depuis 2012 : je suis représentante de l'administration à la Commission Consultative Paritaire des personnels non-titulaires. Là aussi, je suis rentrée dans cette commission à une période très particulière, lors de l'application de la loi "Sauvadet" de 2012 qui a permis aux agents contractuels de la fonction publique de devenir titulaires de leur grade par des recrutements réservés (avec ou sans concours) ou des sélections professionnelles. Ce dispositif a été reconduit jusqu'en 2018.
Depuis 2010 : les premières années j'ai été élue représentante MCF du Conseil de direction du service commun d'Action Sociale (ASUR) dans lequel est notamment voté le budget et les évolutions des actions, ma 2ème casquette de chargée de mission me donnait un rôle particulier que je continue à assumer depuis la restructuration dans le pôle de la DRH, j'avais par exemple insufflé la fête de l'été pour améliorer la cohésion des personnels de l'Université. Nous l'organisons chaque année, en plus de l'arbre de noël. Je participe également mensuellement à la commission Rennes 1 - Rennes 2 d'aides exceptionnelles pour le personnel dont les dossiers nous sont présentés par l'assistante sociale.

Participation aux conseils de composantes, de laboratoires...

Entre 2010 et 2012, j'ai été élue membre du Comité Pédagogique de l'UFR Médecine. Encore une fois, ma participation à cette instance de l'UFR tombe à un moment charnière des études de médecine. En 2010, la PACES (1ère année commune des études en santé) remplace la PCEM1 (1ère année de médecine), la fusion de cette première année notamment des étudiants de médecine, pharmacie et dentaire a nécessité des modifications pédagogiques importantes votées dans ce comité.
Depuis 2019, je suis membre d'un groupe de travail bimensuel composé des directions d'enseignement et de recherche de mathématiques et des collègues présents dans les conseils de Rennes 1, ce groupe aide à la préparation des conseils et à la direction de l'UFR de mathématiques et de l'IRMAR.

Autres

Pour la faculté de Médecine : je suis
- interlocutrice pédagogique de la Direction du Système d'Information depuis 2008, et
- représentante sur les questions de développement durable depuis 2020. L'Université de Rennes 1 est la troisième université à avoir obtenu le label DDRS (en 2019). Les actions initiées dès 2012 au sein de l'Agenda 21 de l'université sont valorisées par ce label, basé sur un référentiel portant sur 5 axes : gouvernance, formation, recherche, gestion environnementale, politique sociale et ancrage territorial.

Pour l'IRMAR : je suis membre de la commission parité et égalité professionnelle entre les hommes et les femmes depuis 2020, nos réflexions et actions couvrent trois axes : les activités scientifiques, le vivre-ensemble et l'éducation. Cela se traduit concrètement par des propositions adressées aux instances dirigeantes.

6.5 Investissement pédagogique

6.5.1 Présentation des enseignements

Mes enseignements sont essentiellement de statistique : analyse descriptive et inférentielle, modèles linéaires et modèles linéaires généralisés, modèle linéaire mixte, modélisation compartimentale en épidémiologie, et de programmation sous R et SAS, en passant par la méthodologie de la recherche clinique et de la méta-analyse. La totalité de mes enseignements a été des créations et non des reprises de cours existants dans mon université. La construction des cours s'est donc faite à partir de différents livres et de cours en ligne, et mes cours de modélisation compartimentale en épidémiologie sont basés sur ma thèse de doctorat et ma recherche actuelle. Le master 2 dans lequel je suis le plus impliquée, est co-habilité sur tout le grand ouest, et j'enseignais donc en visioconférences déjà bien avant la pandémie de COVID 19 (ou plutôt en mode hybride puisque j'ai toujours eu les étudiants rennais dans la même salle que moi, et les autres dans d'autres salles de visio sur Nantes, Tours, Poitiers ou Brest). En 2012, j'ai suivi une formation de 3h sur la classe virtuelle, dispensée par notre service de pédagogie et des TICE. Une autre particularité de ce master est l'origine diversifiée de ses étudiants médecins, pharmaciens, biologistes, statisticiens ou mathématiciens, cette mixité apporte beaucoup aux cours et oblige à une grande réactivité. La particularité de mes enseignements de programmation est une grande hétérogénéité des étudiants avec une obligation à personnalisation de l'enseignement pour les aider à trouver leur propre solution. Je passe donc dans les rangs, tout en gérant le groupe entier dépassant souvent une trentaine d'étudiants, je m'adapte au groupe en proposant parfois des exercices différenciés. J'ai également développé un module en ligne de maîtrise de la rédaction scientifique d'un rapport de projets de Master 2 (avec l'aide d'une ingénieure pédagogique), je m'appuie sur ce module en ligne pendant ma remise à niveau de pré-rentrée de master 2. Mes enseignements depuis 2006 sont détaillés en tableau 6.2.

Nos masters sont essentiellement en formation initiale mais sont ouverts à la formation continue pour la reprise d'études, et chaque année nous avons un ou deux étudiants en formation continue, ce qui enrichit d'ailleurs souvent l'interactivité en cours. En plus de ces enseignements en présentiel en amphithéâtre (promotions de 250 médecins, 65 étudiants de master 1, 120 étudiants de 2ème année ENSAI), en visio-conférences hybrides sur des promotions plus petites de l'ordre de 18 en moyenne mais qui monte en puissance (27 cette année), en salles de travaux dirigés ou en salles informatiques (promotions qui augmentent également 37 au semestre dernier), j'ai eu l'occasion de participer à l'encadrement ou co-encadrement de projets d'ingénieurs de 3 mois à 3 groupes de 3 étudiants de $2^{\text{ème}}$ année ENSAI, et à l'encadrement universitaire d'une stagiaire de Master 2 dans un laboratoire de recherche qui en avait besoin.

Année	Diplôme	Cours + TD	Effectifs	Par an
Depuis 2006*	M2 Master Santé Publique Parcours MPCE	Remise à niveau + Modèle linéaire généralisé	18 par an, soit 300 $\sim$	80 HETD
Depuis 2006*°	M1 Master Santé Publique	Biostatistique + programmation sous R + programmation sous SAS	30 par an, soit 500 $\sim$	100 HETD
Depuis 2017	M1 Master Nutrition et sciences des aliments	Biostatistique + programmation sous R	65 par an, soit 390 $\sim$	32 HETD en 2 groupes
Depuis 2017	M2 Master Ingénierie nutraceutique	Méta-analyse	20 par an, soit 120 $\sim$	6 HETD
Depuis 2009*	4e année Pharmacie	Régression linéaire	100 par an, soit 1300 $\sim$	8 HETD en 4 groupes
Depuis 2012	3e année Médecine	Biomédecine quantitative : modèles compartimentaux	250 par an, soit 2750 $\sim$	3 HETD
Depuis 2005*	3e année Ecole d'ingénieurs ENSAI filière sc. de la vie	Modélisation compartimentale en épidémiologie	18 par an, soit 300 $\sim$	18 HETD
2015-2018*	M2 Master Santé Publique Parcours MTIBH	Régression linéaire	6 par an, soit 12 $\sim$	22 HETD
2018-2020	1re année Santé (Médecine/Pharmacie/...)	Régression linéaire	50 par an, soit 100 $\sim$	4 HETD en 2 groupes
2006-2016	L3 Licence pro SIS à l'IUT de Vannes (UBS)	Méthodologie et statistique pour la recherche clinique	25 par an, soit 275 $\sim$	24 HETD
2015-2016	2e année ENSAI	Modèle linéaire généralisé	120 étudiants	36 HETD
2020	Performances de modèles de prédictions à partir de motifs séquentiels : insuffisance cardiaque, mortalité et trajectoires de soins	Encadrement de projets 2A ENSAI (3 étudiants)		6 séances de 1h sur 3 mois
2018	- Comparaison de biomarqueurs prédictifs dans le cancer du sein. - Evolution temporelle de l'utilisation des méthodes d'analyse de mesures répétées dans les essais thérapeutiques.	Encadrement de projets (2 groupes de 3 étudiants) 2ème année ENSAI		12 HETD
2016	Évaluation des traitements antibiotiques préventifs précoces en médecine vétérinaire, stage de M2 MPCE à l'ANSES	Accompagnement universitaire		4 fois 1h

** sauf 2016-17 où CRCT, et 2008-2009 où congé maternité avec une trentaine d'heures remplacées.*

° sauf 2017-18, 2018- 19, 2020-21 et 2021-22 où j'ai contribué à la formation en enseignement de deux doctorantes en leur confiant une trentaine d'heures de mes cours et TD.

Table 6.2 – Tableau des enseignements

Par ailleurs, outre l'évaluation de mes étudiants sur mes propres enseignements (examens écrits, projets) et plus globalement dans les jurys d'UEs ou d'années dans lesquels j'interviens, je participe à de nombreux jurys d'évaluation de stages, projets ou thèses (tableau 6.3).

Évaluation de travaux de 2 mois à 3 ans	**Nombre d'étudiants**
Examinatrice de master 2 MPCE depuis 2006 (travaux de 6 mois)	(en moyenne 6 par an dont 2/6 à rapporter)
Examinatrice de thèse de sciences (travaux de 3 ans)	2
Membre de jury stage ENSAI (travaux de 6 mois)	1
Présidente de jury projets ENSAI (travaux de 3 mois)	6
Rapporteur de thèse de médecine générale (travaux d'1 an)	1
Membre de comité de suivi de thèse de sciences (travaux de 3 ans)	2

Table 6.3 – Jurys d'évaluation de travaux

En parallèle, j'ai contribué à des opérations de vulgarisation scientifique, en présentant par exemple le métier de biostatisticienne à des lycéennes lors des journées Filles et Maths en 2017 (en plus des tables rondes annuelles d'échanges avec elles pour les encourager à s'orienter vers les maths en post-bac), ou encore pour les hématologues au travers de « focus statistiques » détaillés plus loin dans la partie diffusion du savoir du chapitre recherche.

6.5.2 Responsabilités pédagogiques

Enfin, mon implication dans les diplômes où j'enseigne se poursuit avec la responsabilité pédagogique d'une UE de M2 MPCE (plus celle de remise à niveau), 3 UE de M1 Santé Publique, et d'une UE de M1 NSA. Je suis membre du Comité de direction du Master 2 Modélisation en Pharmacologie Clinique et Epidémiologie (MPCE), et correspondante double cursus des étudiants de 3ème année ENSAI dans le cadre de l'Option Formation Par la Recherche et de la convention ENSAI/M2 MPCE. De 2008 à 2016 j'étais membre du comité de direction et coordinatrice de la Mention Santé Publique (passant de 3 à 6 spécialités, puis de 6 à 10, avec co-habilitation d'Universités du Grand Ouest et de l'EHESP). J'ai pris en charge la co-rédaction (avec construction de formations) de 2 dossiers d'habilitation et 1 dossier d'auto-évaluation, et la responsabilité du M1 Santé Publique du parcours Sciences Médicales (avec entretiens professionnels annuels de la secrétaire pédagogique).

6.5.3 Diffusion, rayonnement, activités internationales

Lors de ma mobilité de 6 mois à l'Université de Southampton, j'ai pu être intégrée dans l'équipe pédagogique de statistique (staff hebdomadaires pédagogiques et de recherche), parfois élargie aux mathématiques, notamment dans leurs moments informels quotidiens (déjeuners, cafés, randonnée annuelle appelée « random walk »), j'ai même pu revêtir l'habit de cérémonie des enseignants pour la remise des diplômes. Nos formations ont de nombreux points communs en statistique et en programmation en R, un montage Erasmus a été évoqué, mis en suspens à cause du Brexit et de la COVID 19. En plus de ma dernière collaboration en recherche, j'ai gardé plusieurs contacts privilégiés là-bas et reprendrai dès que possible des discussions de collaborations pédagogiques d'échanges d'enseignants et d'échanges d'étudiants.

6.6 Glossaire

AG : Andersen-Gill (méthode d'analyse d'évènements récurrents)
AMUE : Agence de Mutualisation des Universités et Etablissements d'enseignement supérieur ou de recherche et de support à l'enseignement supérieur ou à la recherche
ANR : Agence Nationale de la Recherche
AR : AutoRegression
ARA-2 (antagoniste des récepteurs de type 1 de l'angiotensine II
ARGO : AutoRegression with Google search data
ARS : Agence régionale de santé
ASUR : service d'Action Sociale de l'Université de Rennes 1
ATER : Attaché Temporaire d'Enseignement et de Recherche
AVFT : Association européenne contre les Violences faites aux Femmes au Travail
CA : Conseil d'Administration
CAF : Caisse d'Allocation Familiale
CDC : Center for Disease Control (CDC) and Prevention, Organisme de Santé Publique américain
CFIES : Colloque Francophone International sur l'Enseignement de la Statistique
CIC : Centre d'Investigation Clinique
CIFRE : Conventions Industrielles de Formation par la REcherche
CHSCT : Comité d'Hygiène, de Sécurité et des Conditions de Travail
CHU : Centre Hospitalier Universitaire
CNU : Conseil National des Universités (section 85 : Sciences physico-chimiques et ingénierie appliquée à la santé)
CNRS : Centre National de la Recherche Scientifique
CPAM : Caisse Primaire d'Assurance Maladie
CNAMTS : Caisse Nationale d'Assurance Maladie des Travailleurs Salariés
CpG : Régions de densité accrue de séquence dinucléotidiques Cytosine-phosphate-Guanine
CRCT : Congé Recherche et Conversion Thématique
CROUS : Centre Régional des Œuvres Universitaires et Scolaires
DDRS : Développement Durable et Responsabilité Sociétales
DEA : Diplôme d'Etudes Approfondies (Bac+5, prédécesseur des masters recherche)
DGS : Directeur ou Directrice Général(e) des Services
DMSO : DiMéthylSulfOxyde
DRH : Directeur, Directrice ou Direction des Ressources Humaines
EHESP : Ecole des Hautes Etudes en Santé Publique
eHOP : Entrepôt de données biomédicales de l'HOPital de Rennes
ELISA : Enzyme-Linked ImmunoSorbent Assay
EM : Expectation Maximization (algorithme itératif d'optimisation d'une vraisemblance)
ENS : Ecole Normale Supérieure
ENSAI : Ecole Nationale de la Statistique et de l'Analyse de l'Information
ESEN : Ecole Supérieure de l'Education Nationale
FSDIE : Fonds de Solidarité et de Développement des Initiatives Etudiantes
GAN : Generative Adversarial Network (méthode de machine learning)
GPS : Global Positioning System (système de géolocalisation)
HCERES : Haut Conseil de l'Évaluation de la Recherche et de l'Enseignement Supérieur
HETD : Heure Equivalent TD
IHC : ImmunoHistoChemistry
INSERM : Institut National de la Santé Et de la Recherche Médicale
INSHARE : INtegrating and SHAring health data for REsearch

IRMAR : Institut de Recherche MAthématique de Rennes
IA : Intelligence Artificielle
IBS : Integration of the Brier Score
IRA : Infections Respiratoires Aiguës
IRT : Institut de Recherche Technologique (BCom à Rennes sur les technologies numériques)
JSTAR : Journées de la Statistique Rennaises
LASSO : Least Absolute Shrinkage and Selection Operator
LATIM : LAboratoire de Traitement de l'Information Médicale
LRU : loi relative aux Libertés et Responsabilités des Universités
LSTM : Long Short-Term Memory
LTSI : Laboratoire Traitement du Signal et de l'Image
M1 : 1ère année de Master
M2 : 2ème année de Master
MCF : Maître ou Maîtresse de ConFérences
MESRI : Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation
MGEN : Mutuelle Générale de l'Education Nationale
MGMT : gène O6-methylguanine-DNA methyltransferase
MIASHS : Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales
MPCE : Modélisation en Pharmacologie Clinique et Epidémiologie (parcours de master de Santé Publique de l'Université de Rennes 1)
MSE : Mean Square Error
MTIBH : Méthodes de Traitement de l'Information Biomédicale et Hospitalière (parcours de master de Santé Publique de l'Université de Rennes 1)
N+1 : manager direct d'un agent
NSA : Nutrition et Sciences des Aliments (mention de master de Rennes 1)
npde : Normalised Prediction Distributions Errors
OMS : Organisation Mondiale de la Santé
PACES : Première Année Commune aux Etudes de Santé
PAGE : Population Approach Group Europe (organise des congrès spécialisés en analyse de données par approche de population)
PCEM1 : 1ère année du Premier Cycle d'Etudes Médicales
PAS : Prévention, Aide, Suivi
PCC : Pearson Correlation Coefficient
PCR : Polymerase Chain Reaction
PK : pharmacocinétique
PK-PD : pharmacocinétique-pharmacodynamique
PWP : Prentice, William et Peterson (méthode d'analyse d'évènements récurrents)
QVT : Qualité de Vie au Travail
R : logiciel statistique en libre accès (sur le site du Comprehensive R Archive Network)
RF : Random Forest
RH : Ressources Humaines
RMSE : Root Mean Square Error
ROC : Receiver Operating Characteristic
RPS : Risques Psycho-Sociaux
RSF : Random Survival Forest
S3RI : Southampton Statistical Sciences Research Institute
SAEM : Stochastic Approximation Expectation-Maximisation
SAS : Statistical Analysis Software (logiciel de statistique)
SAVE : Service d'Aide à la Vie Etudiante

SFdS : Société Française de Statistique
SIMPPS : Service Inter-universitaire de Médecine Préventive et de Promotion de la Santé
SIR : Susceptible Infectious Removed
SIS : Statistique et Informatique pour la Santé (licence professionnelle)
SNDS : Système National des Données de Santé
SMUT : Service de Médecine Universitaire du Travail
SVM : Support Vector Machine (méthode de machine learning)
SVR : Support Vector Regression
TKI : Inhibiteur de Tyrosine Kinase
TICE : Technologies de l'Information et de la Communication pour l'Enseignement
TD : Travaux Dirigés
TP : Travaux Pratiques
TTE : Time-To-Event
UBS : Université de Bretagne Sud
UFR : Unité de Formation et de Recherche
UE : Unité d'Enseignement
UMR : Unité Mixte de Recherche
UR1 : Université de Rennes 1
UR2 : Université Rennes 2
U.S : United States (Etats Unis)
VP : Vice-Président(e)
VPC : Visual Predictive Check
VEGF : Vascular Endothelial Growth Factor
VSS : Violences Sexistes et Sexuelles
WLW : Wei-Lin-Weissfeld (méthode d'analyse d'évènements récurrents)

6.7 Sélection de 3 articles

Poirier C, **Lavenu A**, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods : Comparison Study. **J Med Internet Res. JMIR Public Health Surveill. 2018** Dec 21 ;4(4) :e11361. doi : 10.2196/11361.
`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6320394/`

Comets E, **Lavenu A**, Lavielle M. Parameter Estimation in Nonlinear Mixed Effect Models Using saemix, an R Implementation of the SAEM Algorithm. **Journal of Statistical Software. 2017** Aug 29. doi : 10.18637/jss.v080.i03.
`https://www.jstatsoft.org/article/view/v080i03`

Quillien V, **Lavenu A**, Karayan-Tapon L, Carpentier C, Labussière M, Lesimple T, Chinot O, Wager M, Honnorat J, Saikali S, Fina F, Sanson M, Figarella-Branger D. Comparative assessment of 5 methods (methylation-specific polymerase chain reaction, MethyLight, pyrosequencing, methylation-sensitive high-resolution melting, and immunohistochemistry) to analyze O6-methylguanine-DNA-methyltranferase in a series of 100 glioblastoma patients. **Cancer. 2012** Sep 1 ;118(17) :4201-11. doi : 10.1002/cncr.27392. Epub 2012 Jan 31.
`https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.27392`

Résumé

Cette Habilitation à Diriger des Recherches (HDR) de Mathématiques s'inscrit en statistique appliquée à l'épidémiologie et à la recherche clinique.

Les technologies modernes permettent de générer des données sur des milliers de variables ou d'observations, que ce soit via internet qui peut être utilisé pour construire un signal médical, ou directement à partir de données-patients génomiques, radiomiques, des bases de données médico-administratives, de surveillance des maladies par des dispositifs médicaux intelligents, etc. Ces données massives nécessitent le développement de méthodes spécifiques statistiques ou d'apprentissage qui rendent robustes et fiables les réponses aux questions cliniques sous-jacentes. Mon travail de recherche s'intéresse principalement aux données de durées (d'apparition d'événements) et de séries temporelles de propagation d'épidémies. Les particularités de ces deux types de données engendrent des sujets méthodologiques qui nécessitent également des études de simulations.

Les modèles présentés, développés ou appliqués sont, entre autres, des forêts aléatoires classiques ou de survie, des méthodes à noyaux adaptées, des modèles autorégressifs, des modèles de Cox, des modèles compartimentaux SIR (Susceptible-Infectious-Removed), des modèles mixtes linéaires et non linéaires, ou des modèles joints.

Ce travail concerne essentiellement des applications médicales d'enjeu de santé publique majeur comme la grippe, la gastro-entérite, la COVID19, le glioblastome, la croissance tumorale, ou les variations de quantité de fer et d'hepcidine pendant le cycle menstruel.

Abstract

This Research Accreditation (HDR) in Mathematics consists of statistics applied to epidemiology and clinical research.

Modern technologies make it possible to generate data on thousands of variables or observations, either via the internet that can be used to build a medical signal, or directly from genomic patient data, radiomics, medico-administrative databases, disease surveillance by smart medical devices, etc. This big data requires the development of specific statistical or machine learning methods in order to provide answers to the underlying clinical questions that are robust and reliable. My research work is mainly concerned with data on durations (time to event) and time series of epidemic spread. The particularities of these two types of data present complex challenges which also require simulation studies.

The models presented, developed or applied are, among others, classical or survival random forests, support vector machines, autoregressive models, Cox models, SIR (Susceptible-Infectious-Removed) compartmental models, linear and nonlinear mixed models, or joined models.

This work mainly concerns medical applications of major public health issues, like influenza, gastroenteritis, COVID19, glioblastoma, tumor growth, or variations of iron and hepcidin during the menstrual cycle.